GENOMIC STUDIES OF RED DRUM (*Sciaenops ocellatus*) IN US WATERS

A Dissertation

by

CHRISTOPHER MICHAEL HOLLENBECK

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	John R. Gold
Co-Chair of Committee,	Carlos F. Gonzalez
Committee Members,	Paul B. Samollow
	James J. Cai
	J. Spencer Johnston
Intercollegiate Faculty Chair,	Dorothy Shippen

May 2016

Major Subject: Genetics

Copyright 2016 Christopher M Hollenbeck

ABSTRACT

Red drum (*Sciaenops ocellatus*) is an economically important marine fish species that supports a large recreational fishery in the United States and is cultured for both restoration and commercial purposes. Characterizing patterns of genetic diversity in wild populations of red drum is essential for understanding how genetic variation in the species is partitioned across space and time and for informed decisions regarding management of the recreational fishery and culture of the species. Advances in DNA sequencing technology now have allowed cost-effective genotyping of thousands of genetic markers and strategies for mapping those markers to the genome. This has led to an unprecedented level of resolution in characterizing patterns of genetic variation in wild populations. The objectives of this research were to supplement the existing red drum linkage map with additional anonymous and gene-linked microsatellite loci, to use the framework provided by the microsatellite-based map to saturate the map with genetic markers (SNP haplotypes) derived from next-generation sequencing, and to use the saturated linkage map, combined with genotypes of wild red drum, to (i) identify potential changes in genetic effective population size over time, and (ii) conduct a population genomic assessment of red drum in U.S. waters. A dense linkage map, consisting of 2,275 genetic markers, was generated. In addition, a method was developed to utilize the linkage map, along with data from studies of linkage disequilibrium, to detect changes in effective population size over time. The method was used to show a recent, temporary decline in effective population size in a sample of red

drum from Matagorda Bay, Texas. A population genomic assessment of red drum revealed three distinct populations of red drum, corresponding to regions in the western Gulf of Mexico, the eastern Gulf of Mexico, and along the southeast Atlantic coast of the U.S. Signatures of natural selection (adaptive variation) were detected among sampled populations, and a set of environmental variables correlated to allele frequencies of loci potentially under selection was identified. Using the linkage map, 15 clusters of loci potentially under the influence of selection were mapped to individual chromosomes, and a set of candidate genes were identified, using comparative genomics. The result of the project is a set of genetic tools and information that will greatly benefit future study of red drum in a variety of contexts.

ACKNOWLEDGEMENTS

For this work, I owe a great debt of gratitude to my advisor, Dr. John Gold, whose enthusiasm for teaching first showed me that the field of genetics offers innumerable problems to solve and that solving them is a worthy pursuit. His advice, mentorship, and support have been an essential component of my professional and personal life. I am also greatly indebted to Dr. David Portnoy, who over the years has endured many hours of questions, and as a result has taught me much of what I know of the practical aspects of research. His guidance has been indispensable to every part of this work.

I have many people to thank for their contribution to this project. In particular, my committee members, Dr. Paul Samollow, Dr. J. Spencer Johnston, Dr. Carlos Gonzalez, and Dr. James Cai, have provided guidance and helpful comments throughout. Dr. Jon Puritz, Dr. Stuart Willis, and Dr. Shannon O'Leary have all contributed significantly through many discussions, help in the laboratory, and with data analysis. I would also like to thank my past and present colleagues at the Marine Genomics Laboratory, including (in no particular order): Dr. Evan Carson, Dr. Trevor Krabbenhoft, Mark Renshaw, Ashley Hanna, Melissa Giresi, Courtney Caster, Dannielle Kulaw, Pavel Dimens, Liz Hunt, Amanda Barker, and Dominic Swift. Robin Waples provided extremely helpful comments on the third chapter of this dissertation.

This work would not have been possible without the work of many collaborators, including researchers at the Florida Fish and Wildlife Conservation Commission, the

South Carolina Department of Natural Resources, the University of Southern Mississippi, the Georgia Department of Natural Resources Coastal Resources Division, and the Texas Parks and Wildlife Department (TPWD). In particular, Robert Vega, Ruben Chavez, and the employees at TPWD's Marine Development Center were essential for generating the mapping crosses and providing tissue samples of red drum brood fish. Research for this project has been supported by Award NA10NMF4270199 from the Saltonstall-Kennedy Program of the National Marine Fisheries Service (National Oceanic and Atmospheric Administration), Award NA10OAR4170099 from the National SeaGrant Program to Texas Sea Grant, a Grants-in-Aid of Graduate Research Award (NA14AR4170102) from Texas Sea Grant, the Coastal Fisheries Division of the Texas Parks and Wildlife Department, Texas AgriLife Research under Project H-6703, and the Harte Research Institute.

Finally, I would like to thank my wife, Allie, whose patience, sacrifice, and love has made all of this possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

BACKGROUND

Red drum, *Sciaenops ocellatus*, is an economically valuable marine fish species supporting both important recreational fisheries and commercial aquaculture in the southeastern United States. The native distribution of the species is throughout the Gulf of Mexico (hereafter Gulf) from Tuxpan, Mexico to southwestern Florida and along the Atlantic coast of the United States (hereafter Atlantic) from southeastern Florida to Massachusetts (Pattillo *et al.*, 1997). Adult red drum, which can live to more than 50 years of age (Ross *et al.*, 1995), typically are found offshore, outside of bays and estuaries, and up to 70 miles from the coast (Davis, 1990). Adults move inshore annually to spawn around the mouths of bays and estuaries, and fertilized eggs are transported by surface currents into bays and estuaries where juveniles remain until sexual maturity at 3-6 years of age (Matlock, 1987; Davis, 1990).

The species at one time supported a large commercial fishery, but in response to rapid declines in red drum abundance beginning in the mid-1980s (Goodyear, 1991), various regulatory measures and harvest restrictions were implemented in the fishery. These include a complete harvest moratorium for red drum in federal waters of the Gulf and Atlantic (GMFMC, 1996; ASMFC, 2002), closure of the commercial fishery in state waters for virtually all Gulf coast and U.S. South Atlantic states, and prohibition of the sale of wild-caught red drum (Matlock, 1990). In addition, red drum stock enhancement

programs have since been implemented in Texas, Florida, Georgia, and South Carolina (McEachron *et al.*, 1995; Woodward 2000; Smith *et al.*, 2001; Tringali *et al.*, 2008).

Today, red drum is arguably the most important recreational marine fish in bays and estuaries in U.S. waters of the Gulf and Atlantic, contributing significantly in 2011 to the >$18 billion in economic impact to (and >100,000 jobs in) Gulf and Atlantic coastal communities from saltwater fishing (Southwick Associates, 2013). Interestingly, the red drum recreational harvest in the Gulf in 2013 totaled ~17.4 million pounds, which is larger than the total harvest of the commercial fishery (~14.1 million pounds) immediately prior to closure of the commercial fishery in the 1980s (NMFS, 2015a). Red drum also is an emerging aquaculture species. Global commercial aquaculture production of red drum has risen sharply in recent years, from 4.4 million pounds in 2000 to 136 million pounds in 2012, largely as a result of increased production in China (FAO, 2015). Despite the fact that red drum is endemic to U.S. coastal waters, a relatively small proportion of global commercial production takes place in the United States; nevertheless, red drum currently is the highest-producing marine fish species in the U.S. by live weight and total sales (USDA, 2014). In 2013, approximately 3.3 million pounds of red drum, valued at over U.S. $10 million, were produced by commercial farms in the U.S. (USDA, 2014).

GENETICS APPROACHES TO FISHERIES MANAGEMENT

One facet of proper management of marine species is characterization of population or 'stock' structure. Characterizing population structure is important because failure to recognize structure within an exploited fishery can lead to over-exploitation and depletion of a localized, undetected stock and loss of unique genetic diversity in that stock (Carvalho and Hauser, 1994; Begg *et al.,* 1999; Hilborn *et al.,* 2003). Loss of genetic diversity can compromise long-term sustainability (Hilborn *et al.,* 2003), and for fisheries undergoing rebuilding, failure to recognize cryptic stocks can result in failure to anticipate patterns of recruitment (Ruzzante *et al.,* 1999).

Population structure of red drum in U.S. waters has been investigated in the past through analysis of both mitochondrial DNA (mtDNA) sequences (Gold *et al.,* 1991, 1993, 1999; Seyoum *et al.,* 2000) and microsatellite loci (Gold and Turner, 2002). These studies showed that red drum exhibit a weak pattern of genetic divergence between the Gulf and Atlantic and that there is an isolation-by-distance effect on allele frequencies across the species' range; a more fine-scale description of population structure, however, has not been forthcoming. A recent study (Michaelsen, 2015), based on18 nuclear-encoded microsatellites and mtDNA control region sequences found no evidence of population structure, based on either marker type, among red drum sampled across the northern Gulf from Texas to Florida. However, these previous studies of red drum have been limited by several factors. Red drum is a large, long-lived marine species with large effective population sizes (N_e) and potentially high levels of gene flow (Carson *et al.,* 2009). This presents a two-fold problem in that genetic heterogeneity

tends to be small and difficult to detect in highly dispersive and abundant populations of marine fishes, and the absence of (detected) genetic heterogeneity does not preclude the existence of independent demographic stocks. The reason for this is that (i) genetic differences at selectively neutral genetic markers such as those used previously take considerable time to accumulate in distinct populations of large N_e, and (ii) fewer migrants are required to homogenize genetic differences between subpopulations at such markers than are needed to remove demographic independence (Waples, 1998; Hauser and Carvalho, 2008). In addition, as it was only possible previously to screen individuals at a few molecular markers, the power to detect weak or recent population structure has been constrained.

GENOMIC APPROACHES

Today, as a result of rapid advances in DNA sequencing technology, a number of techniques have been developed to screen thousands of genetic markers in populations of individuals of virtually any species. One such technique, restriction-associated DNA (RAD) sequencing, uses a next-generation genotyping-by-sequencing approach to identify *de novo* and genotype thousands of single nucleotide polymorphisms (SNPs) (Baird *et al.*, 2008; Davey *et al.*, 2010). This approach offers a solution to the aforementioned problems by: (i) increasing the number of sampled loci, thus increasing the power of detecting small, but significant genetic divergence between populations (caused by very recent isolation) that would be indicative of spatial genetic differences; and (ii) sampling molecular markers that are potentially 'hitchhiking' or proximal to loci

under the influence of natural selection. Because selection can cause allele frequencies to diverge faster than genetic drift alone, particularly in populations with large N_e, identification of loci 'hitchhiking' with genes under differential selection is an alternative method of inferring population structure in recently diverged populations (Bradbury *et al.*, 2010; Bourret *et al.*, 2013).

Genome-wide analyses, while powerful, also adds a level of complexity to population genetics analysis and inferences. Computing population statistics with thousands of loci simultaneously is computationally intensive and identification of genomic outliers among large numbers of markers presents additional problems, including an elevated risk of identifying false positives. In this regard, it is especially useful to have information regarding the genomic context for each locus. A locus that is observed to be under the influence of selection, for example, is less likely to be a false positive if tightly-linked markers also exhibit the same pattern of differentiation. Further, given the possibility that observed markers show signals of selection due to genomic 'hitchhiking', a comparative genomic approach can be used to identify potential nearby candidate genes that may be the actual targets of selection. While there is not currently a draft genome assembly available for red drum, a genetic linkage map containing the SNP markers to be analyzed in the population study is a suitable alternative. In addition to providing positional information for markers, linkage maps have the advantage of providing estimates of recombination rate between genetic markers, which can be combined with linkage disequilibrium data to provide estimates of current and past N_e of populations (Hill, 1981; Hayes *et al.*, 2003).

PROJECT OVERVIEW

The overall goal of this research is to thoroughly explore the ways in which next-generation DNA sequencing can be used to better understand genetic variation and divergence among wild populations of red drum. Specifically, the objectives of the research are to: (i) supplement the second-generation red drum genetic map with additional neutral and gene-based microsatellite loci; (ii) use the framework provided by the microsatellite-based linkage map to create a SNP-based linkage map, using next-generation sequencing technology; (iii) use information in the genetic map to estimate past and present N_e of wild red drum populations; and (iv) conduct a genomics-based evaluation of population structure and investigate the extent and organization of neutral and potentially adaptive genetic variation in red drum in U.S. waters of both the Gulf and Atlantic.

CHAPTER II

A MICROSATELLITE-BASED GENETIC LINKAGE MAP OF RED DRUM (*Sciaenops ocellatus*) AND COMPARISON OF CHROMOSOMAL SYNTENIES WITH FOUR OTHER FISH SPECIES*

INTRODUCTION

A central problem in commercial aquaculture is maximizing production efficiency. Genetic improvement of farmed aquatic species has been suggested as a permanent and cumulative solution to this problem (Gjedrem *et al.*, 2012). Most traits targeted by selective breeding programs are influenced by many genes (quantitative trait loci, QTLs) with cumulative effects and/or more complex interactions (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Genetic marker-based breeding schemes that exploit linkage associations between easily screened genetic markers and QTLs offer advantages over traditional breeding programs, particularly for traits that are difficult to measure and for species with relatively long generation times (Hulata, 2001; Sonesson, 2007). Genetic linkage maps of polymorphic markers are a critical first step in establishing marker-based selection programs and also provide a framework for physical mapping and genome assembly (Liu and Cordes, 2004; Danzmann and Gharbi, 2007).

---

*Reprinted from *Aquaculture* volume 435, C. M. Hollenbeck, D. S. Portnoy, and J. R. Gold, "A microsatellite-based genetic linkage map of red drum (*Sciaenops ocellatus*) and comparison of chromosomal syntenies with four other fish species", pp. 265-274, Copyright 2015 with permission from Elsevier.

An alternative strategy for identifying QTLs is a candidate gene approach where *a priori* information about a gene's biological function is used to predict that gene's impact on a trait of interest (Lynch and Walsh 1998). This approach has been used in fishes to identify QTLs affecting spawning time (Leder *et al.,* 2006), growth rate (Tao and Boulding, 2003), and sex determination (Shirak *et al.*, 2006). Further, the advent of next-generation DNA sequencing has led to the generation of massive amounts of genetic sequence data for many fish species, including a whole genome assembly of the economically important Nile tilapia (*Oreochromis niloticus*). The ever-increasing availability of DNA sequence data facilitates candidate gene approaches through comparative genomics by taking advantage of interspecies synteny – the possession of similar chromosomal regions due to common descent – to transfer relevant genomic information obtained from studies on well-characterized species to studies involving emerging species (Sarropoulou *et al.,* 2007). One way of identifying synteny between species is to assess the distribution of shared genetic markers in both genomes and identify regions where a common ordering of those markers occurs. Type-I (protein-coding) genetic markers (O'Brien, 1991) are ideal for this approach as they are often conserved between species, and when incorporated into a linkage map can provide a framework for comparative genomics analysis.

Here, we present a genetic linkage map for red drum, expanding upon previous work (Karlsson *et al.,* 2007; Portnoy *et al.,* 2010, 2011) by the addition of 177 anonymous microsatellites and 46 microsatellites closely linked to Type-I loci. We report the map locations of a total of 486 microsatellites, including the 46 linked to

Type-I loci, spanning all 24 (haploid) red drum chromosomes. We also demonstrate the application of the genetic map as a tool for candidate-gene identification through comparative genomics by putatively localizing an additional 80 known (but previously unmapped) red drum coding genes and microsatellites closely linked to Type-I loci (EST-SSRs), using a synteny-based mapping approach.

MATERIALS AND METHODS

In previous studies (Karlsson *et al.,* 2007; Portnoy *et al.,* 2010, 2011), two full-sib mapping families (Family A, n=103; and Family B, n=104), generated from outbred, single-pair crosses carried out at the Marine Development Center of the Texas Parks and Wildlife Department (TPWD), were used. This study took advantage of the same tissue samples used in those studies; details of crosses, spawning, egg collection, and larval grow-out may be found in Portnoy *et al.,* (2010) and references therein.

A total of 177 polymorphic, anonymous microsatellites were isolated from a repeat-enriched library. Details of enriched-library preparation, primer sequences, and summary statistics for each microsatellite can be found in Renshaw *et al.,* (2012). In addition, 133 expressed sequence tag-linked microsatellites (EST-SSRs) were designed following the comparative approach outlined in Hollenbeck *et al.,* (2012). Summary information, including repeat motif, primer sequences, and putative identity for all EST-SSRs can be found in Hollenbeck *et al.,* (2015). Genomic DNA was extracted following a modified Chelex extraction protocol (Estoup *et al.,* 1996). Following removal of residual Chelex by centrifugation at 16,000 x g, one microliter of supernatant was used for each PCR reaction. The 177 anonymous microsatellites and the 46 microsatellites

linked to Type-I loci (= 223 total) yielded mapping-informative genotypes in at least one parent and were subsequently genotyped in the appropriate progeny. Genotyping was conducted following procedures outlined in Portnoy *et al.*, (2010).

Because individuals genotyped in this study also were used in prior mapping efforts, genotype data from the 223 microsatellites were combined with genotypes at the 264 microsatellites assayed previously by Karlsson *et al.,* (2007) and Portnoy *et al.,* (2010, 2011). Linkage analysis was conducted with the program JOINMAP v4.1 and linkage groups were defined initially by using microsatellites previously assigned to the 24 red drum linkage groups (Portnoy *et al.,* 2011). New markers were assigned to existing groups, using an LOD threshold of 3.0. Marker order for each linkage group was computed using the maximum-likelihood (ML) mapping function implemented in JOINMAP. Tests for segregation distortion for each marker were carried out using a chi-square goodness-of-fit test; probabilities of individual genotypes, conditional upon the map order, were computed to check for possible genotyping errors. A preliminary map was generated for each parent and marker order was compared between individuals to ensure order agreement. If marker order for each linkage group was in agreement across all parents, a family-specific map was generated using the multipoint ML algorithm for map construction with full-sib outbred families, as implemented in JOINMAP and described in van Ooijen (2011). Briefly, the algorithm generates separate ML maps for each parent in a cross and integrates the maps by averaging distances between shared intervals and interpolating or extrapolating positions of markers segregating in only one of the parents. Family-specific maps were then checked for marker order agreement.

Finally, both family-specific maps were integrated into a consensus map, using the program MERGEMAP (Wu *et al.*, 2011).

The consensus map was used to compare the red drum genome with assembled genome sequences and chromosome designations of four other fishes: Nile tilapia (*Oreochromis niloticus*), stickleback (*Gasterosteus aculeatus*), green spotted puffer (*Tetraodon nigroviridis*), and fugu (*Takifugu subripes*). The most recent assembly of each species' genome (tilapia, v 1.1; stickleback, v 1.0; green spotted puffer, v 8; fugu, v 5) was downloaded to a Linux server. The discontiguous-megablast algorithm in NCBI's BLAST+ suite (Camacho *et al.*, 2009) was used to compare flanking sequences of the original clone of each mapped red drum microsatellite or EST sequence (for EST-SSRs) with each comparison genome. Clone sequences were available on GENBANK for 429 of the 440 mapped anonymous microsatellites; clone sequences of 11 of the anonymous microsatellites were not available. In total, 475 mapped microsatellites (429 anonymous and 46 EST-SSRs) were used in the BLAST search. Matches were considered similar if they had a region of ≥ 50 base pairs of overlap and had an e-value of $\leq 10^{-10}$. To prevent duplicated sequences from confounding results, only sequences with a single match within a genome were considered for further analysis. Chromosome number and chromosomal position (in base pairs) was recorded for each hit, and Oxford plots comparing the red drum linkage map to the genome of each of the four comparison species were generated, using the GRID graphics package in R (Murrell 2005). As the comparison species most relevant to aquaculture, tilapia was chosen for a more detailed analysis of synteny with red drum. To visualize the extent of marker collinearity

between red drum and tilapia, positions of markers with significant matches to the tilapia genome were coded as relative positions along the length of their respective chromosomes/linkage groups. For tilapia, the start position of the marker, in base pairs, was divided by the total length of the chromosome, in base pairs. For red drum, the position of each marker, in centiMorgans (cM), on the consensus map was divided by the total length of the linkage group, in cM. Based on the observation that the majority of individual red drum linkage groups corresponded to individual tilapia chromosomes, the latter were reorganized along the y-axis such that chromosomes homologous between the two species aligned along the diagonal axis of the graph. Shared markers were then plotted based on their relative positions on linkage groups/chromosomes. A custom Perl script (available upon request from CMH) was used to identify blocks of shared synteny between red drum chromosomes and chromosomes of each of the four comparison species. Syntenic blocks were defined as sets of markers on the same linkage group and in the same order in both species, uninterrupted by any other shared marker. Ordering mismatches between markers that were separated by less than five percent of the total length of a linkage group/chromosome were ignored in order to maximize detection of informative syntenies otherwise disrupted by small-scale, local rearrangements or ordering errors caused by uncertainty in the mapping process. Syntenic regions from chromosomes involved in apparent chromosomal rearrangements between species detected from the Oxford plots were plotted as circular ideograms, using the software CIRCOS v0.66 (Krzywinski *et al.*, 2009).

A synteny-based mapping approach was used to identify likely locations of red drum coding genes that were stored on NCBI's GENBANK and of EST-SSRs that were designed by Hollenbeck *et al.,* (2012) but which were monomorphic in mapping families and could therefore not be mapped via linkage analysis. Nucleotide sequences for 85 red drum coding genes were downloaded from GenBank and reduced to 72 novel nucleotide sequences by excluding duplicates of the same locus. These sequences and the 87 monomorphic EST-SSRs were compared by BLAST search to each of the four comparison genomes, using the same criteria mentioned above. Given that these loci are known to exist in the red drum genome, a locus that maps in another species to a syntenic region shared between red drum and that species likely exists in the same region of the red drum genome. Thus, these loci were mapped to the genomes of the four comparison species, and when red drum genes and EST-SSRs mapped into computed syntenic regions in at least one other species, the locus was putatively localized to that marker interval in red drum.

RESULTS

The map for Family A contained 372 microsatellites, including 32 linked to Type-I loci; the map for Family B contained 406 microsatellites, including 34 linked to Type-I loci. The map for Family A had a total size of 1641.2 cM, with an average linkage group size of 68.38 cM and an average marker interval of 4.81 cM; the map for Family B had a total size of 1722.0 cM, with an average linkage group size of 71.75 cM and an average marker interval of 4.55 cM. The consensus map (Figure 2.1) contained 486

microsatellites, including 46 linked to Type-1 loci. The total size, average linkage group size, and average marker interval of the consensus map was 1815.3 cM, 75.64 cM, and 3.96 cM, respectively. A single microsatellite, *Soc*685, which was mapped to linkage group 8 in a previous study (Portnoy *et al.*, 2010), was removed from the final map due to the presence of significant segregation distortion in all four parents. Of the 46 microsatellites linked to Type-1 loci, 38 (82.6%) could be assigned a putative identity following a BLASTN search of NCBI's nucleotide (nt) database.

Of the 429 anonymous microsatellites with available clone sequences, 163 (38.0%) had significant homology to the tilapia genome. Of these, six were excluded from further analysis due to hits on multiple chromosomes. A total of 41 (89.1%) of the 46 microsatellites linked to Type I loci produced a significant hit to the tilapia genome; of these, three were excluded due to hits on multiple chromosomes. The total number of BLAST hits across both microsatellite types was 204 (tilapia), 154 (stickleback), 105 (fugu), and 84 (green spotted puffer).

Oxford plots for all species (Figure 2.2) revealed significant homology between red drum linkage groups and chromosomes of the four comparison species, with an approximate one-to-one relationship observed between red drum linkage groups and the chromosomes of each of the species. A number of both intra- and inter-chromosomal rearrangements, however, appear to have occurred since red drum and each of the four comparison species diverged from a common ancestor. Examples of inferred

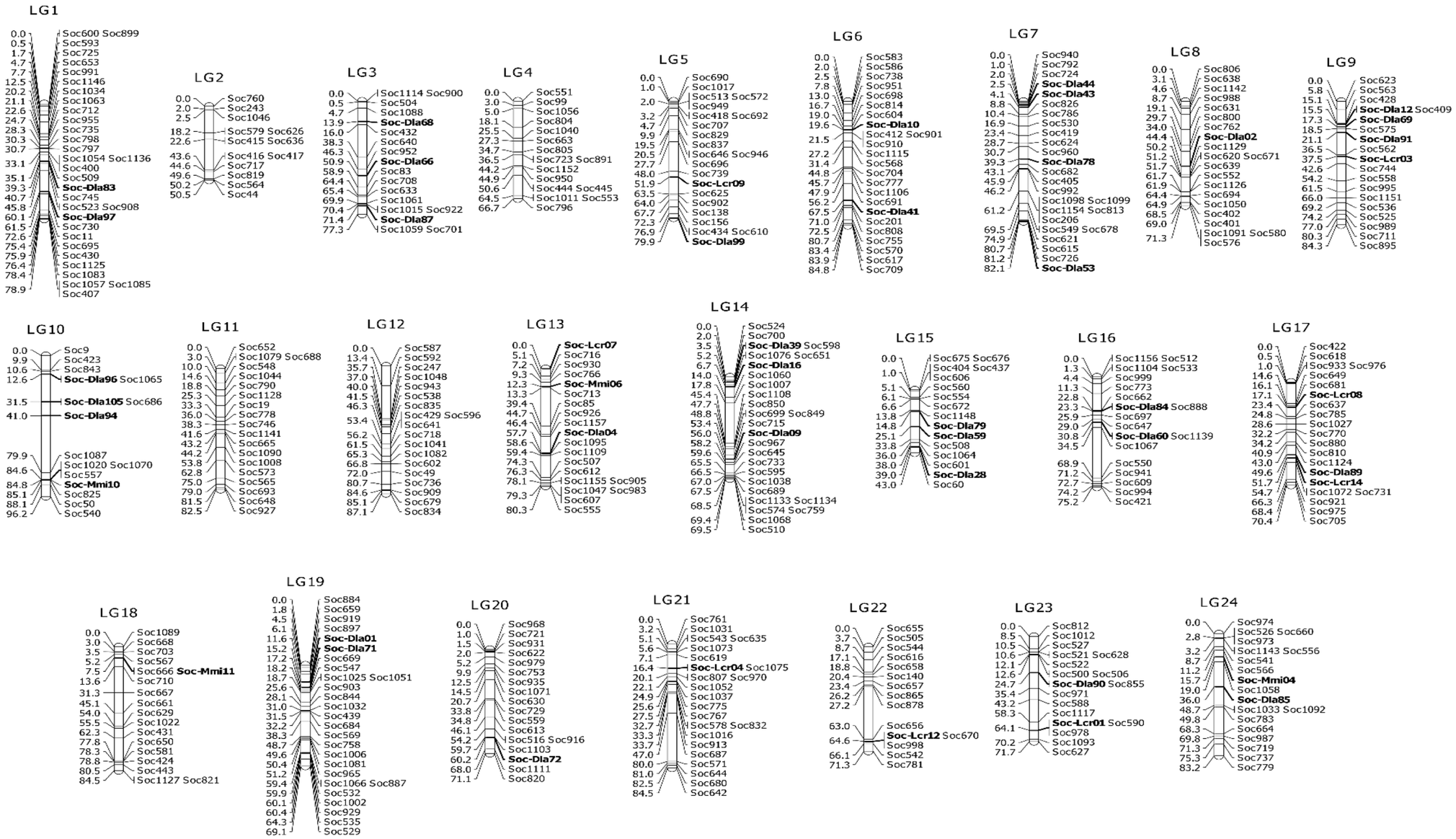


Figure 2.1 Consensus microsatellite linkage map. The map is based on segregation in two full-sib families of red drum, *Sciaenops ocellatus*. Map distances, in cM, are given to the left of each linkage group (LG), while marker names are given on the right; marker names in bold represent gene-linked (Type-I) microsatellites. *Reprinted with permission from Hollenbeck et al., (2015).*

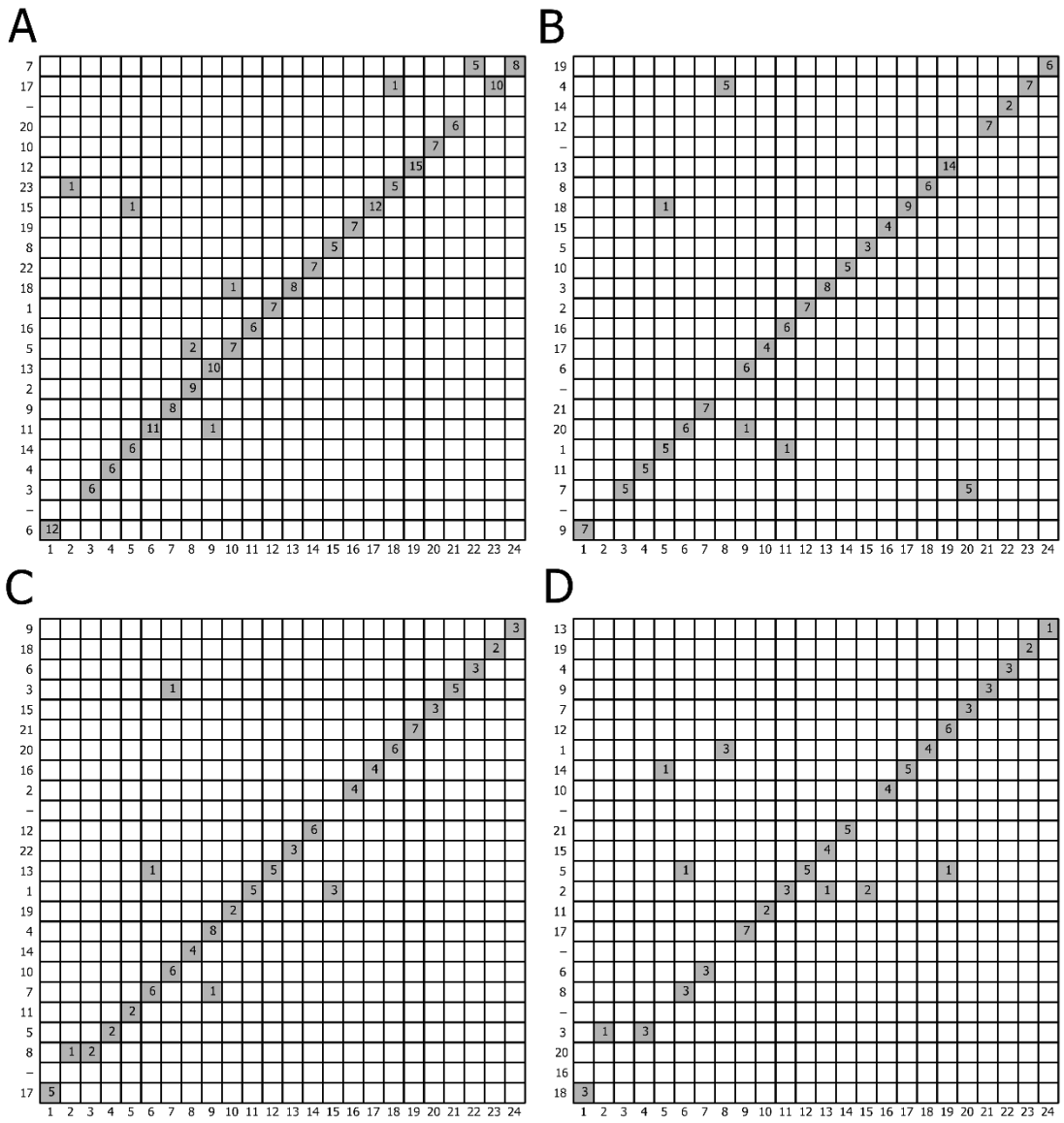


Figure 2.2 Oxford plots displaying conserved synteny. Plots are between linkage groups of red drum and chromosomes of four comparison species. Abscissa: linkage groups 1-24 of red drum; ordinate: chromosomes of comparison species, arranged by homology to linkage groups of red drum. Comparison species are: A, Nile tilapia (*Oreochromis niloticus*); B, three-spined stickleback (*Gasterosteus aculeatus)*; C, Japanese pufferfish (*Takifugu rubripes)*; and D, green spotted pufferfish (*Tetraodon nigroviridis)*. Numbers in grid squares indicate the number of markers (loci) shared between a red drum linkage group and a chromosome in a comparison species. *Reprinted with permission from Hollenbeck et al., (2015).*

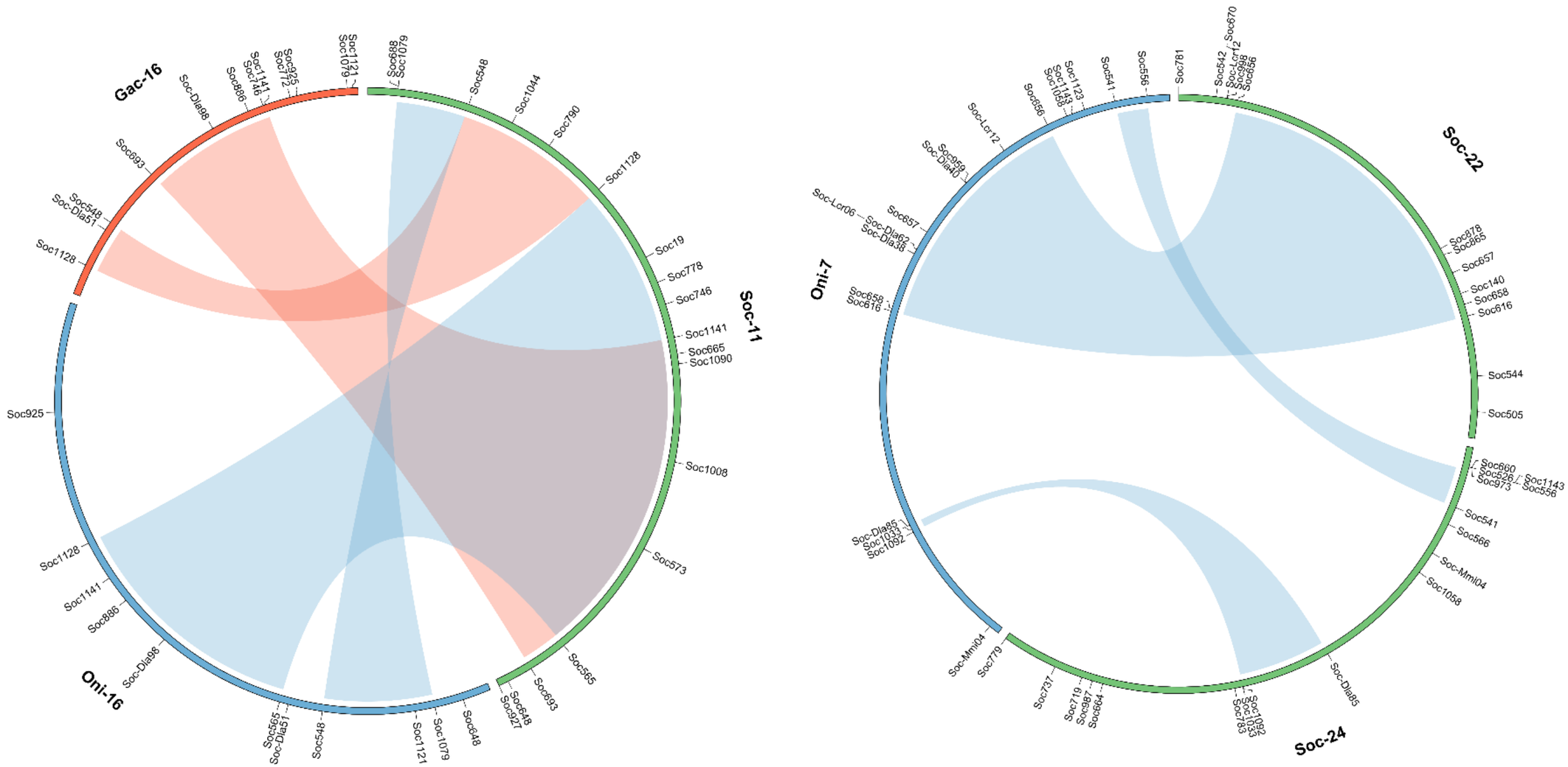


Figure 2.3 Circular ideograms. Plots were generated using CIRCOS v0.66 and showing putative intra- and inter- chromosomal rearrangements that occurred since red drum and Nile tilapia diverged from a common ancestor. Left, an inferred chromosomal rearrangement involving syntenic groups on red drum linkage group 11 (*Soc*-11) and Nile tilapia chromosome 16 (*Oni*-16). Ribbons linking chromosomes represent regions of shared synteny. The two syntenic regions derived from *Soc*-11 are inverted on *Oni*-16 relative to their position on *Soc*-11. Asterisks indicate the ends of relocated syntenic blocks. Right, an inferred fusion involving syntenic groups on two red drum chromosomes (*Soc*-22 and *Soc*-24) occurring on a single Nile tilapia chromosome (*Oni*-7). A fusion in the Nile tilapia lineage is inferred because *Soc*-22 and *Soc*-24 are syntenic to separate chromosomes in the other three comparison species (data not shown). The two syntenic regions from *Soc*-24 flank a single syntenic region from *Soc*-22 on *Oni*-7, suggesting that a fusion was followed by one or more intra-chromosomal rearrangements. *Reprinted with permission from Hollenbeck et al., (2015).*

chromosomal rearrangements, based on shared syntenic group locations, are presented in Figure 2.3. Several instances where linkage groups on the same red drum chromosome occurred on more than one chromosome of a comparison species were observed; these included five instances in tilapia and three instances each in the other three comparison species. Finally, a comparison of marker order between putatively homologous chromosomes in red drum and tilapia revealed large regions of synteny and shared marker order between the two species (Figure 2.4). Shared markers generally aligned along the diagonal axis of the plot, which is expected if markers are largely collinear.

Based on criteria described above, 47 conserved syntenic regions were identified between red drum and tilapia and a total of 172 microsatellites were placed into these regions. The number of microsatellites per shared syntenic region ranged from two to ten, with a mean of 3.66. Combined, syntenic regions spanned 306 Mb (46.6%) of the tilapia genome assembly and 838.29 cM (46.2%) of the red drum map. In addition, 33, 30, and 23 syntenic regions were identified between red drum and stickleback, fugu, and green spotted puffer, respectively; syntenic regions spanned 37.8% (stickleback), 36.5% (fugu), and 32.1% (green spotted puffer) of the species' genome assemblies.

Of the 72 coding genes in red drum available on GENBANK, 50 had a single hit to the genome of at least one of the four comparison species. Of these, 28 (50.6%) were mapped to a genomic interval by synteny-based mapping. Of the 87 monomorphic EST-SSRs in red drum, 79 had a single hit to the genome of at least one of the four comparison species; 52 of these (65.8%) were mapped with the same approach. Fifty of the EST-SSRs were assigned a putative identity based on a BLASTN search of NCBI's nt

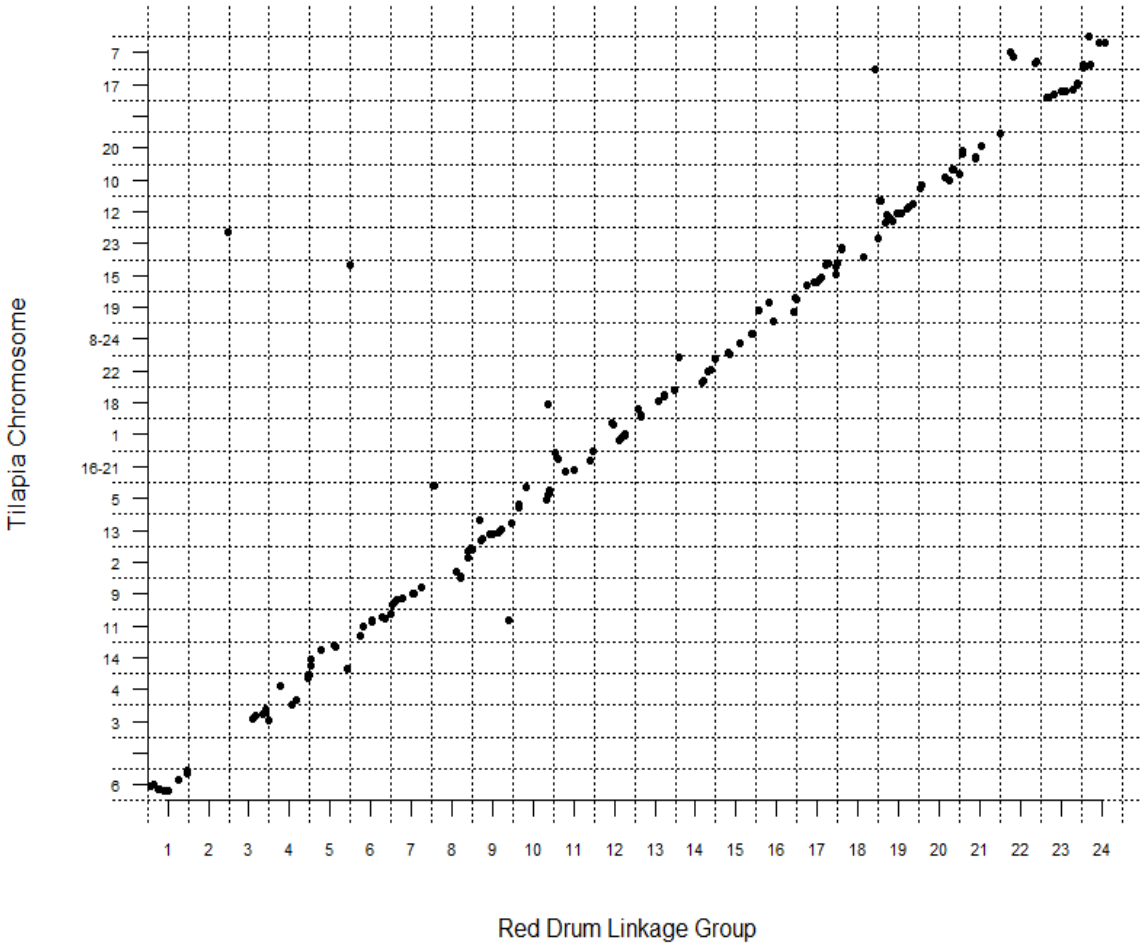


Figure 2.4 Comparison of synteny between red drum and Nile tilapia. Abscissa: linkage groups 1-24 of red drum; ordinate: Nile tilapia chromosomes, arranged by homology, to linkage groups of red drum. Shared markers are plotted relative to their position on a given chromosome/linkage group. *Reprinted with permission from Hollenbeck et al., (2015).*

database. A summary of the 28 coding genes and the 52 EST-SSRs, including GENBANK accession number, putative identity, flanking markers, species in which the syntenic regions are conserved, linkage group in red drum, and interval size, is given in Table 2.1. The map locations of 17 of the coding genes and of 25 of the EST-SSRs were supported by shared synteny in more than one of the four comparison species.

DISCUSSION

An additional 227 microsatellites were added to the red drum map, increasing the total number of mapped microsatellites to 486 (440 anonymous, 46 linked to Type-I loci). The addition of these microsatellites decreased the inter-marker interval from 6.28 cM (previous sex-average map) to 3.96 cM. The total length of the consensus map was 1815.3 cM. This is larger than the size (1196.9 cM) of the sex-averaged map reported previously (Portnoy *et al.,* 2010), for two possible reasons. First, the additional microsatellites sampled more of the chromosomal content of the red drum genome by mapping locations distal to markers on the previous map; and second, a component of the difference is likely attributable to the process of merging family-based maps into a single consensus map. Because of known differences in recombination rate between sexes in red drum (Portnoy *et al.,* 2010) and differences in marker polymorphism between individual parents, distances between markers in the consensus map may only be reflective of that of a single individual. If a marker is only segregating in one sex, marker intervals involving that locus in the consensus map will not be a sex-averaged distance, but will reflect only the recombination rate in that particular sex (which could be larger than the sex-average). In addition, while the software MERGEMAP outperforms

Table 2.1 Summary of synteny-mapped loci. Accession No. – GENBANK accession number of a gene sequence or EST; Function – the assigned gene name from GENBANK or a significant BLASTN hit (ESTs); Flanking Loci – the closest red drum markers between which the locus from GENBANK could be mapped based on synteny with another species; Comparison Species – the species in which a syntenic relationship with red drum existed: 1: three-spined stickleback, 2: green spotted puffer, 3: Nile tilapia, 4: fugu; Linkage Group – the red drum linkage group to which the locus was mapped; Interval – the size of the corresponding marker interval in centiMorgans on the red drum map. *Reprinted with permission from Hollenbeck et al., (2015).*

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
AF062520.1	*Sciaenops ocellatus* somatolactin precursor	*Soc*646	*Soc*418	1,3	5	17.28
AF064872.1	*Sciaenops ocellatus* translation initiation factor eIF-2B precursor	*Soc*810	*Soc*880	1,2,3,4	17	6.68
AY677170.1	*Sciaenops ocellatus* salmon-type gonadotropin-releasing hormone precursor	*Soc-Lcr*03	*Soc-Dla*91	1,3	9	16.4
AY677171.1	*Sciaenops ocellatus* chicken II-type gonadotropin-releasing hormone precursor	*Soc-Mmi*10	*Soc*1065	1	10	72.2
AY876899.1	*Sciaenops ocellatus* hemoglobin beta chain	*Soc*1148	*Soc-Dla*28	4	15	25.22
FJ415100.1	*Sciaenops ocellatus* peptidoglycan recognition protein II	*Soc-Dla*97	*Soc*1125	4	1	16.33
GQ384067.1	*Sciaenops ocellatus* 11 beta-hydroxylase (CYP11B)	*Soc-Dla*10	*Soc*1115	1,3	6	7.62
GQ384068.1	*Sciaenops ocellatus* 21-hydroxylase (CYP21)	*Soc-Dla*10	*Soc*1115	3	6	7.62

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FJ641038.1	*Sciaenops ocellatus* neuronal nitric oxide synthase	*Soc-Dla*01	*Soc-Dla*71	2,3,4	19	3.66
GU144512.1	*Sciaenops ocellatus* glycoprotein alpha subunit	*Soc*1139	*Soc*550	2,3	16	38.15
GU144513.1	*Sciaenops ocellatus* thyrotropin beta subunit	*Soc*1087	*Soc-Dla*96	1,2	10	67.26
GU799603.1	*Sciaenops ocellatus* insulin-like growth factor I	*Soc*978	*Soc-Dla*90	2,4	23	39.33
GU368832.1	*Sciaenops ocellatus* recombination activating protein 1 (RAG1)	*Soc*719	*Soc*1092	1	24	22.63
GU368812.1	*Sciaenops ocellatus* si:dkey-174m14.3 gene	*Soc*810	*Soc*1072	4	17	13.86
GU370888.1	*Sciaenops ocellatus* ISG15	*Soc*758	*Soc*569	1,3,4	19	10.37
GU929942.1	*Sciaenops ocellatus* viperin (Vip)	*Soc*1139	*Soc*550	2,3	16	38.15
HM581689.1	*Sciaenops ocellatus* putative tissue factor pathway inhibitor 1	*Soc*1141	*Soc*1128	4	11	16.33
HM368401.1	*Sciaenops ocellatus* putative tissue factor pathway inhibitor 2 (TFPI2)	*Soc-Dla*09	*Soc*1108	1	14	10.57
HQ651238.1	*Sciaenops ocellatus* high mobility group protein B1 (HMGB1)	*Soc*949	*Soc*1017	3,4	5	0.98

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
HQ731135.1	*Sciaenops ocellatus* FIC domain-containing protein (ficd)	*Soc-Dla*01	*Soc-Dla*71	1,2,3,4	19	3.66
HQ731297.1	*Sciaenops ocellatus* receptor-interacting serine-threonine kinase 4 (RIPK4)	*Soc-Dla*72	*Soc*820	1,3	20	10.92
JX002675.1	*Sciaenops ocellatus* eukaryotic translation initiation factor 3 subunit G (eTIF3)	*Soc*1040	*Soc*804	2,4	4	7.42
JX002676.1	*Sciaenops ocellatus* NADH dehydrogenase 1 alpha (ND1)	*Soc*1141	*Soc*1128	4	11	16.33
JQ938122.1	*Sciaenops ocellatus* hypothetical protein (GCS1)	*Soc*991	*Soc*1063	2	1	13.46
JQ938817.1	*Sciaenops ocellatus* peroxisomal enoyl-CoA hydratase/L-3-hydroxyacyl-CoA dehydrogenase (EHHADH)	*Soc*565	*Soc*1141	1,3,4	11	33.45
JQ939810.1	*Sciaenops ocellatus* LOC562320 (KIAA1239)	*Soc-Dla*68	*Soc*640	1,4	3	24.44
KC830168.1	*Sciaenops ocellatus* Sushi/von Willebrand factor type A/EGF/pentraxin domain-containing 1 (SVEP1)	*Soc-Dla*68	*Soc*640	1	3	24.44
KF140446.1	*Sciaenops ocellatus* T-box brain 1 (tbr1)	*Soc*565	*Soc*1141	1,3,4	11	33.45

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FP242838.1	*Oreochromis niloticus* arginine-glutamic acid dipeptide (RE) repeats (rere)	*Soc*1052	*Soc*578	1,2,4	21	10.62
FM010232.1	*Neolamprologus brichardi* calcium/calmodulin-dependent protein kinase type II subunit gamma-like (LOC102781265)	*Soc-Dla*59	*Soc-Dla*79	1,3	15	10.3
FK943099.1	*Anoplopoma fimbria* Beta-synuclein	*Soc-Dla*02	*Soc*1129	1,2,3,4	8	5.85
FP237559.1	*Neolamprologus brichardi* guanine nucleotide-binding protein G(s) subunit alpha-like (LOC102788485)	*Soc-Mmi*10	*Soc*1065	1	10	72.2
FP241017.1	*Maylandia zebra* DNA damage-binding protein 1-like (LOC101485195)	*Soc*718	*Soc*1048	4	12	19.2
FP238020.1	*Oreochromis niloticus* protein phosphatase 1 regulatory subunit 14B-like (LOC100711861)	*Soc-Dla*66	*Soc*708	1,3	3	13.45
FL488459.1	*Oreochromis niloticus* protein-L-isoaspartate(D-aspartate) O-methyltransferase-like (LOC100708432)	*Soc-Lcr*14	*Soc-Dla*89	3,4	17	2.11

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FK941535.1	*Oreochromis niloticus* elongation of very long chain fatty acids protein 6-like (LOC100706271)	*Soc*694	*Soc*1050	3	8	0.48
CV186185.1	*Danio rerio* si:dkey-11e23.5 (si:dkey-11e23.5)	*Soc*810	*Soc*1072	4	17	13.86
FM028201.1	*Oreochromis niloticus* phosphatidylserine synthase 1 (ptdss1)	*Soc-Dla*10	*Soc*412	1,3,4	6	1.92
FM001773.1	*Oreochromis niloticus* ATPase asna1-like (LOC100702925)	*Soc*991	*Soc*1063	2,3	1	13.46
FK940504.1	*Haplochromis burtoni* protein FAM212A-like (LOC102290510)	*Soc*825	*Soc-Mmi*10	1,3,4	10	0.25
FM000143.1	*Maylandia zebra* prospero homeobox protein 1-like (LOC101476671)	*Soc*810	*Soc*880	1,2,3,4	17	6.68
FM023318.1	*Oreochromis niloticus* thioredoxin reductase 3 (txnrd3)	*Soc*423	*Soc-Dla*96	1	10	2.69
FM027384.1	*Neolamprologus brichardi* AF4/FMR2 family member 4-like (LOC102796839)	*Soc*630	*Soc*1071	2,4	20	6.16

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FM004496.1	*Oreochromis niloticus* disco-interacting protein 2 homolog B-A-like (LOC100703539)	*Soc*1052	*Soc*578	1,2,4	21	10.62
FM009184.1	*Neolamprologus brichardi* hippocalcin-like protein 1-like (LOC102792629)	*Soc-Dla*09	*Soc*849	3	14	7.16
FM011451.1	*Haplochromis burtoni* myristoylated alanine-rich C-kinase substrate-like (LOC102308505)	*Soc*975	*Soc*921	3	17	2.07
FM000141.1	*Oreochromis niloticus* junction plakoglobin-like (LOC100707214)	*Soc*430	*Soc-Dla*97	1,3	1	15.84
FK940790.1	*Oreochromis niloticus* serine-rich coiled-coil domain-containing protein 2-like (LOC100710988)	*Soc-Lcr*03	*Soc-Dla*91	1	9	16.4
FM010695.1	*Neolamprologus brichardi* neurotrypsin-like (LOC102780776)	*Soc-Dla*59	*Soc*1148	3	15	11.33
FM012479.1	*Neolamprologus brichardi* MAP7 domain-containing protein 1-like (LOC102788526)	*Soc*1133	*Soc*645	3	14	8.82
AM986102.1	*Oreochromis niloticus* protein bicaudal D homolog 2-like (LOC100712336)	*Soc*642	*Soc*687	3	21	37.51

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FM000541.1	*Haplochromis burtoni* notch-regulated ankyrin repeat-containing protein A-like (LOC102310701)	*Soc*657	*Soc*658	3	22	4.65
FP238879.1	*Haplochromis burtoni* mucin-17-like (LOC102294623)	*Soc-Lcr*12	*Soc*657	1,3	22	41.12
FN565801.1	*Oreochromis niloticus* CCAAT/enhancer-binding protein beta-like (LOC100689715)	*Soc*642	*Soc*687	3	21	37.51
FM013092.1	*Pundamilia nyererei* lysophospholipid acyltransferase 5-like (LOC102193954)	*Soc*1115	*Soc*777	1,3	6	18.5
AM987101.1	*Oreochromis niloticus* LIM domain-binding protein 3-like (LOC100707522)	*Soc-Dla*59	*Soc-Dla*79	1,3	15	10.3
FP237257.1	*Haplochromis burtoni* arf-GAP with dual PH domain-containing protein 1-like (LOC102311544)	*Soc-Dla*79	*Soc-Dla*28	2,4	15	24.19
FM018821.1	*Oreochromis niloticus* ubiquitin specific peptidase 9, X-linked (usp9x), transcript variant X7	*Soc*1128	*Soc*548	1	11	15.24
FP242802.1	*Oreochromis niloticus* calsyntenin-3-like (LOC100707828)	*Soc*1115	*Soc*777	1,3	6	18.5

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FM006783.1	*Oreochromis niloticus* cAMP-dependent protein kinase type I-alpha regulatory subunit-like (LOC100711171)	*Soc-Dla*28	*Soc*601	3	15	0.99
FM017384.1	*Oreochromis niloticus* beta-14-galactosyltransferase 5-like (LOC100696774)	*Soc*578	*Soc*687	2	21	14.27
AM985524.1	*Neolamprologus brichardi* dedicator of cytokinesis protein 8-like (LOC102793453)	*Soc-Dla*71	*Soc*439	4	19	16.31
FM007039.1	*Oreochromis niloticus* zinc finger CCCH domain-containing protein 7B-like (LOC100698451)	*Soc-Dla*79	*Soc-Dla*28	2,4	15	24.19
FM007045.1	*Haplochromis burtoni* breakpoint cluster region protein-like (LOC102292921)	*Soc*657	*Soc*658	3	22	4.65
FM012644.1	*Neolamprologus brichardi* forkhead box protein O3-like (LOC102783221)	*Soc*1072	*Soc*921	4	17	11.57
FM012811.1	*Oreochromis niloticus* beta-14-galactosyltransferase 5-like (LOC100696774)	*Soc*578	*Soc*687	2	21	14.27
FM021649.1	*Neolamprologus brichardi* E3 ubiquitin-protein ligase MSL2-like (LOC102797866)	*Soc*1139	*Soc*550	2,3	16	38.15
FM008475.1		*Soc*1117	*Soc*588	3	23	15.03

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
FM013106.1	*Neolamprologus brichardi* calcipressin-3-like (LOC102786199)	*Soc*1139	*Soc*550	3	16	38.15
FM000627.1	*Haplochromis burtoni* nuclear factor 1 X-type-like (LOC102293245)	*Soc*991	*Soc*1063	2	1	13.46
AM984068.1	*Oreochromis niloticus* cordon-bleu protein-like 1-like (LOC100702457)	*Soc*565	*Soc*1141	1,3,4	11	33.45
CX348550.1	*Oryzias latipes* basic leucine zipper transcriptional factor ATF-like (LOC101168259)	*Soc*810	*Soc*880	1,2,3,4	17	6.68
CX348556.1	*Oreochromis niloticus* basic leucine zipper transcriptional factor ATF-like (LOC100690329)	*Soc*810	*Soc*880	1,2,3,4	17	6.68
C48612.1	*Haplochromis burtoni* breakpoint cluster region protein-like (LOC102292921)	*Soc*657	*Soc*658	3	22	4.65
EV413959.1	*Morone saxatilis* clone apoa1_3 apolipoprotein A-I (ApoA1)	*Soc-Lcr*09	*Soc*646	3	5	31.36
GW668767.1		*Soc*588	*Soc*971	3	23	7.79

Table 2.1 continued

Accession No.	Putative Function	Flanking Loci		Comparison Species	Linkage Group	Interval
GW668773.1	*Maylandia zebra* basic leucine zipper transcriptional factor ATF-like (LOC101468837)	*Soc*810	*Soc*880	1,2,3,4	17	6.68
GW670899.1	*Oreochromis niloticus* nuclear factor erythroid 2-related factor 1-like (LOC100705427)	*Soc*850	*Soc*1108	1,3,4	14	2.28
GW671772.1	*Neolamprologus brichardi* nuclear receptor subfamily 2 group F member 6-like (LOC102781642)	*Soc*1095	*Soc*507	1,3,4	13	15.69
GW672302.1	*Epinephelus coioides* CCAAT/enhancer-binding protein beta 2	*Soc*642	*Soc*687	3	21	37.51

JOINMAP in estimating a merged marker order (Galeano *et al.*, 2011), it also inflates inter-marker distance in combining maps (Khan *et al.*, 2012). However, there exists a tradeoff between accurate estimation of map distances and combining incomplete information from multiple individuals into a single map. For purposes of conserved synteny analysis, establishing the linear order of the maximum number of loci is more useful than having more accurate map distances.

The addition of 46 microsatellites linked to Type-I loci to the red drum map is important, as a large percentage of mapped Type-I loci (82.6%) were assigned a putative function. Further, an appreciably larger percentage of microsatellites linked to Type-I loci, relative to anonymous microsatellites (89.1% vs. 38.0%) were conserved between red drum and tilapia, demonstrating the utility of microsatellites linked to Type-I loci for comparative genomics analysis. A comparison of the red drum linkage map to the genomes of four different percomorph fishes revealed significant conserved synteny and that numerous chromosomal rearrangements had occurred since red drum and each of the comparison species last shared a common ancestor. While a one-to-one chromosomal relationship generally was observed between red drum linkage groups and chromosomes of each of the four comparison species, there were several instances where regions from different red drum chromosomes were found on a single chromosome of a comparison species, and there were several inferred intra- and inter-chromosomal rearrangements. Overall, the findings are consistent with previous comparative genomics studies in teleost fishes where instances of chromosomal repatterning, as well

as a large degree of conserved synteny, have been observed (Kucuktas *et al.*, 2009; Sarropoulou *et al.*, 2007).

The comparatively high degree of synteny conservation, in terms of total number of orthologous syntenic regions, number of loci present in those regions, and percent of genome assembly covered by syntenic regions, between red drum and tilapia was not unexpected. The red drum EST-SSRs were designed by using a comparative approach that utilized an unassembled version of the tilapia genome to ensure maximum cross-species amplification (Hollenbeck *et al.*, 2012), and red drum (Family Sciaenidae; Order Perciformes) and tilapia (Family Cichlidae; Order Perciformes) have been assumed traditionally to be closer phylogenetically than red drum is to either sticklebacks (Order Gasterosteiformes) or fugu and green spotted puffer (Order Tetraodontiformes) (Nelson 2006). The high degree of synteny and conservation of marker order between red drum and tilapia may be useful in future genetic selection of red drum. Tilapias (*Oreochromis* spp.) have been the subject of considerable genetics research related to aquaculture, and QTLs influencing production-relevant traits (e.g., growth rate, immune and stress response, sex determination, cold tolerance) have been identified (Cnaani *et al.*, 2003, 2004; Lee *et al.*, 2003, 2004; Moen *et al.*, 2004; Shirak *et al.*, 2006). Based on the current set of shared loci between the two species, 47 syntenic blocks spanning 306 Mb (46.6% of the tilapia genome assembly) were identified and represent chromosomal regions that have remained intact over evolutionary time and likely share a significant proportion of homologous genes. Further work to identify these genes in red drum will soon be underway.

Using the syntenic regions identified from all comparisons, it was possible to putatively localize an additional 28 red drum coding genes, downloaded from GENBANK, and 52 unmapped EST-SSRs to marker intervals on the red drum map. The 28 coding genes taken from GenBank are of interest as some appear to be involved in immune response. These include: (i) the neuronal nitric oxide synthase gene (nNOS), which is expressed in a number of tissues and thought to be involved in innate immune response (Zhou *et al.*, 2009); (ii) the high mobility group protein B1 (HMGB1), which is up-regulated in response to bacterial challenge and is thought to be involved in immune function (Zhao *et al.*, 2011); and (iii) the product of the tissue factor pathway inhibitor 2 gene (TFPI-2), which is thought to play a role in the response to bacterial infection (Zhang *et al.*, 2011). In addition, 50 of 52 synteny-mapped EST-SSRs were able to be assigned a putative function following a BLASTN search. These include a gene coding for a thioredoxin reductase protein, which has been shown to be expressed during pathogen infection in rainbow trout (Pacitti *et al.*, 2014), and a gene coding for a junction plakoglobin gene product, which has been observed to be upregulated in channel catfish skin tissue in response to pathogen challenge (Li *et al.*, 2013).

In summary, the current linkage map of 486 total microsatellites (440 anonymous, 46 gene-linked) is a powerful tool for comparative genomics. Using synteny-based mapping, we putatively localized an additional 28 red drum coding genes and 52 red drum EST-SSRs. The mapping of highly conserved anchor loci will provide a framework for additional future comparative work and should allow researchers to leverage relevant genomic information from studies involving well-characterized species

to inform candidate-gene approaches to QTL detection in red drum. The general strategy presented for mapping by synteny can be applied to any species without an available genome assembly, but with an available linkage map. In addition, the map potentially will facilitate identification of chromosomal regions under the influence of natural selection in wild populations of red drum, and in this way could inform both management of wild stocks and stock-enhancement decisions. Moreover, the map will be a valuable resource for future genomics research in red drum, including physical mapping and genome assembly.

CHAPTER III

DETECTION OF RECENT CHANGES IN EFFECTIVE POPULATION SIZE FROM LINKAGE DISEQUILIBRIUM BETWEEN LINKED AND UNLINKED LOCI

INTRODUCTION

Measurement of linkage disequilibrium (LD) between pairs of unlinked genetic markers has become the most prevalent method to estimate contemporary effective population size (N_e) in the fields of population and conservation genetics. This is due largely to the relative ease with which the approach can be applied, as it only requires a single sample and ~20 polymorphic genetic markers (Waples, 2006). In addition, well-established analytical methods and software packages for application are available (Waples, 2006, Waples and Do, 2008; Do *et al.*, 2014). While microsatellite loci previously have been the most commonly used genetic markers for applying the LD method, genomics techniques now allow the generation of datasets with genotypes at thousands to tens of thousands of single nucleotide polymorphisms (SNPs). This is beneficial for application of LD-based methods to estimate N_e as the ability to genotype hundreds or thousands of SNPs permits greatly improved precision (Waples and Do, 2010). However, the ability to generate genotypes at many loci distributed across the genome presents a problem in that many of the markers are likely to be linked physically, and if all loci are assumed to be unlinked, estimates of N_e may be downwardly biased due to excess LD caused by linkage rather than drift (Sved *et al.*, 2013). A straightforward solution to this problem is to remove pairwise comparisons involving known linked loci (Larson *et al.*, 2014).

This approach, however, does not take full advantage of all information present in a large SNP dataset. As noted by Hill (1981), while LD from unlinked loci reflects current contemporary N_e (hereafter current N_e), LD from physically linked loci reflects contemporary N_e in past generations (hereafter past N_e). Thus, if information pertaining to physical linkage is available for large number of markers, e.g., in the form of a genome sequence or genetic map, LD can be evaluated across a spectrum of linkage values to remove the downward bias on current N_e caused by linked loci and, in addition, identify potential changes in N_e in prior generations.

Use of LD and linkage data to estimate past N_e largely has been limited to model species because of the need for linkage or genomic position data. Hayes *et al.*, (2003) introduced a novel measure of LD, chromosome segment homozygosity (CSH), which was used in simulated data sets to track changes in N_e over time and, with empirical data, to infer demographic population histories in dairy cattle and humans. Using CSH, they (*Ibid*) also derived an approximate relationship between the degree of linkage (the recombination rate, c) and the number of generations in the past (t) to which an estimate of N_e would apply: $t = \frac{1}{2c}$. Tenesa *et al.*, (2007) expanded upon this by instead using the LD statistic r^2, which has the same expected relationship to N_e as CSH. The authors used r^2 estimated from haplotypes of ~1,000,000 SNPs identified in the human HapMap project (International HapMap Consortium, 2003) to infer a recent increase in human N_e over the last 1,000 years. Subsequently, several studies involving domesticated animals (Corbin *et al.,* 2010; Flury *et al.,* 2010; Qanbari *et al.*, 2010; Alam *et al.,* 2012; Herrero-Medrano *et al.,* 2013) have shown that with extremely dense genotype and genome-

sequence data, estimates of contemporary N_e can be obtained from roughly the previous generation to t generations in the past. However, these studies utilized haplotype-based methods to estimate LD, which require that marker phase is either known or estimable from the data. Haplotype-based estimators require relatively rare, long haplotypes to estimate N_e in the very recent past ($t \leq 50$), and because precision of the estimate is dependent upon the number of locus pairs used (Hill, 1981), estimates representing the recent past are less precise than estimates from the more distant past (Hayes *et al.*, 2003).

There is potential to apply a linkage-based approach to non-model species for which large SNP datasets and linkage information (from linkage maps or whole genomes) are increasingly available. However, marker densities in these species may be relatively low and phased haplotypes cannot be computed with accuracy, meaning that the approach has limited utility for non-model species when investigating processes that act on evolutionary time scales. For example, a linkage map constructed with 100 individuals will only be able to resolve LD at loci separated by 0.01 Morgans (M). Assuming the approximate relationship between recombination rate and time derived by Hayes *et al.*, (2003), this would reflect N_e approximately 50 generations in the past. However, understanding changes in N_e in the recent past (≤ 50 generations) is of great interest to conservation biologists because detecting recent declines (e.g., due to anthropogenic effects) or expansions (due to recovery efforts) are important components of genetic monitoring programs (Luikart *et al.*, 2010). Using a linkage-based approach would have an advantage over traditional LD- (Waples and Do, 2008) and variance-

based (Nei and Tajima, 1981, Pollak, 1983) approaches to detect recent changes in N_e in that it requires only a single genetic sample rather than sampling over multiple years before and after a demographic change (Antao *et al.*, 2010).

Here, we extend the LD-based approach of Waples and Do (2008) by including linkage information to estimate N_e over a range of time points in the past. The advantage of this approach (hereafter, the linkage approach) over haplotype-based methods is two-fold: (i) a composite LD measure is used, enabling calculation of pairwise LD from genotype data in the absence of phase information; and (ii) because the vast majority of locus pairs in the genome are unlinked, high precision for estimating current N_e can be achieved without biases associated with inclusion of physically linked loci. We apply this approach to simulated data to assess the ability to detect demographic changes (changes in N_e) in past generations across a variety of demographic models, using a dataset of 1,000 SNP loci. We also explore issues important to interpretation of the results. These include the importance of correcting for bias caused by small sample size relative to the true N_e, the effect of rare alleles on estimates made at multiple points in time, and the effect of time of sampling relative to a change in N_e. In addition, we compare estimates of current N_e in which physical linkage is taken into account with estimates, based on the same data, where all locus pairs are assumed to be unlinked, in order to quantify bias. Finally, to demonstrate the effectiveness of the method on an actual dataset, we apply the linkage approach to an empirical dataset of SNP genotypes from a sample of a marine fish where a recent, temporary reduction in N_e was known to have occurred.

MATERIALS AND METHODS

The method presented here requires genotype data from a diploid species and a matrix of pairwise recombination rates for all genotyped markers. The latter could be obtained from linkage mapping data or estimated from genome sequence data. The general strategy involves binning estimates of LD between pairs of loci based on similar observed recombination rates (c). Previous work (Hayes *et al.*, 2003) showed that the time period to which an LD-based estimate of N_e applies is a function of c ($t = \frac{1}{2c}$). This equation suggests that time and recombination rate do not scale linearly and that most of the range of possible recombination rates (0 – 0.5 M) relates to generations in the recent past. Thus, bins were defined by generations rather than by recombination rate and calculated as $\frac{1}{2c}$. For these analyses, bins were defined from 1 to 3.33 generations in the past (c = 0.5 to 0.15 M), 3.33 to 5 generations (c = 0.15 to 0.1 M), 5 to 10 generations (c = 0.1 to 0.05 M), and 10 or more generations (c = 0.05 to 0.0 M). We note that bins could be otherwise defined to suit particular research questions. For each bin, weighted estimates of total r^2 (r^2_{total}) and r^2 attributable to sampling variation (r^2_{sample}) for all pairs of loci were obtained, following Waples and Do (2008). The difference between r^2_{total} and r^2_{sample}, which is equal to the component of the total r^2 attributable to genetic drift (r^2_{drift}), and the mean c value of pairs of loci in each bin (in Morgans) were then used to calculate N_e, following Hill (1981) and Waples (2006). A software program, LINKNE, was written in the Perl programming language to facilitate analyses. A detailed description of the program and calculations can be found in Appendix 1.

Simulation

Precision and Bias

Simulations were used to evaluate the effectiveness of the linkage approach to detect changes in N_e under a variety of demographic models and to explore important properties of the method. Simulations were written in Python, utilizing libraries from the program SIMUPOP (Peng and Kimmel, 2005). All simulations included a single, closed 'constant' population with discrete generations, equal sex ratio, and binomially distributed reproductive success, such that the census size (N) is approximately equal to N_e. While N_e under the simulated conditions is actually slightly larger than N, i.e., $N_e = N + 1/(2N) + 0.5$ (Balloux, 2004), the correction term $1/(2N) + 0.5$ was used for all calculations as the 'true' N_e, although for simplicity N and N_e will be treated as equivalent hereafter, following Waples (2006). Populations with an initial (starting) effective size of N_e = 100, 250, 500, and 1000 were used in simulations, with each simulation replicated 100 times. The genome used in simulations consisted of 25 chromosomes, each 0.75 M in size; for each chromosome, map positions for 200 SNP loci were chosen randomly at the beginning of each simulation. Initial allele frequencies at each SNP locus were determined by a pseudo-random draw from a uniform distribution (0, 1). Consequently, each replicate began with loci near linkage equilibrium. Theoretical results (Sved, 1971) indicate that for populations with $N_e \leq$ 1000, loci separated by at least 0.01 M and starting in linkage equilibrium should reach steady-state levels of linkage disequilibrium in approximately 200 generations. Thus, all replicates were 'burned-in' for 200 generations. The per locus mutation rate followed a

SNP-model, with the rate of forward mutation equal to 1×10^{-8} and the reverse mutation rate equal to 1×10^{-9}. The probability of recombination between adjacent loci was proportional to the distance between them, i.e., loci 0.01 M apart have a 1% chance of recombination in each individual in each generation. After 250 generations, 50 individuals were sampled from each simulated population and genotypes at 1000 randomly selected, polymorphic SNP loci were recorded into a single GENEPOP file. A square matrix of recombination rates for all pairs of loci also was generated for each simulated population. For each population, N_e was estimated by using pairs of loci binned as noted previously; loci with minor alleles at frequency < 0.05 were excluded from the analysis. Initial runs revealed a downward bias in estimates of N_e in prior generations due to tightly linked loci that had not reached steady-state linkage disequilibrium; consequently, locus pairs separated by less than 0.015 M were excluded from estimations. Estimates of the coefficient of variation (CV) of N_e, calculated as in Hill (1981), were used to generate 95% confidence intervals for each bin, and harmonic means of estimates of N_e and their confidence intervals across replicates were plotted using the *ggplot2* package (Wickham, 2009) in R (R Core Team, 2015). Bias of each estimate was computed as the distance of the harmonic mean of estimated N_e, across replicates, from the true N_e and expressed as a percentage of the true N_e. Precision was measured as the CV of N_e (Hill, 1981).

Detection of Changes in N_e

Five different demographic models were simulated in addition to the 'constant' population described above. Three models involved declines in effective size (to 25%,

50%, and 75% of starting size), while two involved expansions (to 2x and 5x of starting size). All models were simulated as an instantaneous change in census size that occurred five generations prior to sampling; otherwise, all simulations were run in exactly the same manner as with the constant population. A sample (*S*) of 50 individuals was taken at the end of each simulation except when a model involved a reduction in census size to less than 50 individuals, in which case all remaining individuals were sampled. Detection of a change in N_e was assessed by observing whether confidence intervals overlapped between the estimate of N_e from the most recent bin (1 to 3.3 generations in the past) and the estimate from bin furthest in the past (≥10 generations). Bias and precision of each estimate were evaluated as discussed at the end of the prior section.

Evaluation of Sample-Size Bias Correction

Because the linkage approach is intended to identify possible demographic changes by evaluating differences in N_e, measured using pairs of markers that have various linkage relationships, it is important to determine whether estimates of N_e made from any single bin are more or less biased than estimates from other bins. If so, different levels of bias among bins could be incorrectly interpreted as demographic change. Waples (2006) and England *et al.,* (2006) reported a bias in estimating N_e due to exclusion of second- and higher-order terms when accounting for the contribution of sampling error to LD measured in a finite sample. The bias is downward and is particularly large when *S* is small relative to the true N_e. To account for the bias, an empirically derived correction factor was proposed by Waples (2006). To explore the effect of the correction on estimates of N_e from prior generations, all simulations of the constant population model

were evaluated with both the sample-size bias correction and using only 1/*S* to account for r^2_{sample}. N_e was again measured across 100 replicates and the harmonic mean of results across replicates recorded.

Allele Frequency Cutoff

The presence of rare alleles also can bias LD-based estimates of N_e (Waples, 2006), and excluding rare alleles has been proposed (Waples and Do, 2008) as a means to reduce the bias. We explored the effect of this bias by testing a series of allele-frequency cutoff thresholds (0.10, 0.05, 0.02, 0.01, and 0), using the constant population model with N = 250. Here, and in all subsequent analyses, the sample-size bias correction proposed by Waples (2006) was applied to estimations of N_e. As above, N_e was measured across 100 replicates and results averaged across replicates.

Effect of Time between Demographic Change and Sampling

Over time, drift and recombination reorganize patterns of LD, removing signatures of past N_e. In order to evaluate effectiveness of the linkage approach to detect past demographic change, the length of time that signatures of past N_e persist in the genome was assessed under two different models: a decline of 25% and an expansion to 2x, both with a starting N_e of 250. The simulation was modified to adjust the number of generations (1, 5, 10, 20, and 50) prior to sampling in which the change occurred. As above, results were averaged across 100 replicates.

Comparison to the LDNE Method

More often than not, linkage relationships of marker pairs are not known and current N_e is estimated by LD under the assumption that all loci are unlinked. This assumption becomes compromised when genotypes at thousands of genetic markers are obtained, an increasingly common standard in genomics studies of non-model species (Allendorf *et al.*, 2010). To evaluate the effect of this assumption, we used NEESTIMATOR v.2.01 (Do *et al.*, 2014) to estimate current N_e from the simulated data, using the constant population model for each 'true' N_e (100, 250, 500, and 1000). Estimates of N_e and parametric confidence intervals were obtained by excluding alleles with frequency < 0.05 and recording the harmonic mean from 100 replicates. The difference between estimates of N_e and the true N_e was estimated and compared to results when using only unlinked pairs of loci and LINKNE.

Empirical Data

The linkage approach also was applied to genotype data from a single sample of juveniles of red drum (*Sciaenops ocellatus*) sampled from West Matagorda Bay, Texas, in 2008. West Matagorda Bay is one of several Texas bays and estuaries that are stocked annually with fingerling red drum as part of a state-wide stock enhancement program (Vega *et al.*, 2003) and was one of several bays sampled over a period of years to monitor the relative contribution of stocked fish to wild populations (Karlsson *et al.*, 2008; Carson *et al.*, 2014). The sample from West Matagorda Bay was selected for analysis because it contained an abnormally high proportion (>16%) of juvenile fish of hatchery origin (Carson *et al.*, 2014). Because the hatchery-raised individuals likely

Table 3.1 Estimates of current effective population size for juvenile red drum sampled from West Matagorda Bay, TX. Current effective population size (N_e) was estimated for: (i) all juvenile red drum; (ii) wild individuals only; and (iii) hatchery-raised individuals only. Estimates were generated using NeEstimator2 (Do *et al.*, 2014); low/high refer to parametric 95% confidence intervals. Rare alleles were excluded below a threshold of 0.05. *S* refers to sample size.

Sample	Low	N_e	High
All (S = 56)	206.9	208.3	209.7
Wild (S = 42)	4753.5	5912.6	7817.1
Hatchery (S = 14)	12.4	12.4	12.5

originated from a limited number of breeders (Gold *et al.,* 2008; Carson *et al.,* 2014), the result was a relatively small N_e in the sample that contained both hatchery-raised and 'wild' fish (Table 3.1). In addition, because the reduced N_e is a first-generation effect caused by the presence of a large proportion of hatchery-raised individuals in the sample, the reduction in N_e should not be detected in prior generations.

Population Samples

Tissue samples from 64 juvenile red drum sampled from West Matagorda Bay, TX, in the spring of 2008 were randomly chosen from a larger set of samples collected for a prior study; sample collection procedures were described in Carson *et al.,* (2014). DNA was extracted using a phenol-chloroform-isoamyl (PCI) extraction protocol and double-digest restriction-site associated DNA (ddRAD) libraries were prepared following procedures outlined in Peterson *et al.,* (2012). RAD libraries were sequenced on an Illumina HiSeq 2000 DNA sequencer.

Sequence data were first demultiplexed by individual barcode sequence, using the process_radtags program from the STACKS software package (Catchen *et al.,* 2011). Read mapping and SNP calling were performed with the *dDocent* pipeline (Puritz *et al.,* 2014). Reads were mapped to a reduced-representation genome sequence, developed as part of an ongoing study and consisting of red drum RAD fragments sequenced on the Illumina MiSeq platform and assembled to recover the entire sequence of each fragment. The raw SNP dataset (consisting of SNP, indel, and complex polymorphisms, but hereafter referred to as SNPs, unless otherwise specified) consisted of 524,657 SNP

genotypes for 64 individuals. Genotypes and individuals were filtered based on numerous criteria. First, initial filtering, to remove loci with more than 50% missing genotypes and individuals with less than 10x mean coverage across loci, removed 73,123 SNPs and six individuals from the analysis. Genotypes with a depth of < 10 reads and a Phred quality score of less than 20 were then filtered, followed by removal of loci with more than 5% missing data across individuals and a minor allele frequency of less than 0.05. A total of 6748 SNPs across 3224 RAD-tags remained after these filtering steps. These genotypes were subject to another round of filtering for quality control: filters for allele balance across forward and reverse reads, improper pairing of reads, excessively high depth and low quality score (genotypes), and maximum mean depth (loci). This resulted in 6048 SNPs across 3048 RAD-tags. SNP loci were then filtered further if they did not conform to expectations of Hardy-Weinberg equilibrium (HWE) (*P* value < 0.001); 212 SNPs (66 RAD-tags) were removed. The remaining complex polymorphisms were decomposed to allelic primitives with the vcfallelicprimitives program in the vcflib package (https://github.com/ekg/vcflib), resulting in 6,189 SNPs and indel polymorphisms. These were collapsed into haplotypes for each RAD-tag, using a custom Perl script. The script filtered indel polymorphisms (unless they were the only polymorphism on the RAD-tag) and RAD-tags for which less than 95% of individuals could be successfully haplotyped across the RAD-tag. A GENEPOP file, consisting of (haplotype) genotypes at 2101 multi-allelic RAD-tags was produced using the script and implemented for NE2 and LINKNE analyses.

Mapping Panel

DNA from tissue samples for two parents and 108 full-sib progeny was extracted and ddRAD libraries prepared as above. Samples were obtained as part of prior studies in which a microsatellite-based linkage map was constructed for red drum; details regarding the mapping cross and sample collection can be found in Portnoy *et al.,* (2010) and references therein.

Bioinformatic processing of SNP data was performed as above, with the exception that filters for HWE were not applied to the data. SNPs were collapsed into haplotypes for each RAD-tag and the resulting genotypes output into JOINMAP (van Ooijen, 2012) format. The pairwise recombination rate for each pair of loci was calculated for each parent, using a custom Perl script. When a recombination rate could be estimated for both parents (both loci were segregating in an informative manner), the rate was averaged between parents. When the recombination rate could be estimated for only one parent, the sex-specific rate was used. Recombination rates were output as a square matrix of pairwise values.

Genotypic data and a matrix of pairwise recombination rates were used to generate estimates of N_e, using LINKNE. The program was run as in the simulations except that no filter was applied to remove tightly linked locus pairs. Data were summarized using the ggplot2 package in R.

RESULTS

Simulation

Precision and Bias

Simulations involving populations of constant size were used to assess precision and bias associated with estimates of N_e at different points in the past. The strategy used to bin locus pairs resulted in four bins, each producing an estimate of N_e at a different time in the past. The exact time point used for each estimate (t) was dependent upon the distribution of loci in the genome, which was randomly determined at the beginning of each simulation. Mean time estimates for the four bins, averaged across all starting values of N_e, were 1.01, 3.99, 6.65, and 14.35 generations in the past.

Bias, as measured by the distance between the harmonic mean of N_e estimates across replicates and the true N_e, scaled by true N_e, was less than 10% in all cases; direction and magnitude of the bias was dependent upon both N_e and number of generations in the past to which an estimate applied. Bias for estimates from the most distant past (14.35 generations) was smallest and positive (upward bias) for N_e = 100 (2.29%) and negative (downward bias) for larger N_e (-0.35%, -4.85%, and -4.99% for N_e = 250, 500, and 1000, respectively). Bias for estimates from the most recent past (1.01 generations) was positive (3.54%, 3.15%, 7.51%, and 6.96% for N_e = 100, 250, 500, and 1000, respectively), while bias for intermediate time points in the past (3.99 and 6.65 generations) ranged from -2.42% to -9.26%. In all but one case, confidence intervals for estimates of N_e encompassed the true N_e. Due to a slight upward bias and high precision,

the estimate of N_e from the most recent past (1.01 generations) for the simulation where N_e = 100 had a confidence interval of 100.6 – 107.5.

Precision was greatest for estimates from the more recent past (1.01 generations) and ranged from 0.017 (N_e of 100) to 0.081 (N_e of 1000). The next highest level of precision was obtained for estimates from the most distant past (14.35 generations) and ranged from 0.053 (N_e of 100) to 0.096 (N_e of 1000). Intermediate time points (3.99 and 6.65 generations) were the least precise, ranging from 0.054 (t = 6.65 generations; N_e of 100) to 0.620 (t = 3.99 generations; N_e of 1000).

Detection of Changes in N_e

Results of simulations to investigate the ability of the linkage approach to detect declines and expansions in N_e are summarized in Figure 3.1. For the models where N_e remained constant, confidence intervals always overlapped; thus, a change in N_e was never falsely detected. A change in N_e was detected in 80% of all decline/expansion models where a change in N_e had occurred. Changes in N_e were detected more often when initial effective population size was small and/or when the magnitude of change was great. This was due in part to greater precision of estimates of N_e in smaller populations. A summary of demographic models and whether a change in N_e was detected in each model is presented in Table 3.2.

Estimates of N_e over time for constant and decline models are shown in Figure 3.1a. The linkage approach was always able to detect declines to 25% of initial N_e. Declines to 50% of initial N_e of were detected for initial N_e of 100, 250, and 500 but not 1000; declines to 75% were only detected for initial N_e of 100 and 250. Estimates of N_e

Table 3.2 Sensitivity of detection of changes in N_e for different demographic models. Change in N_e was estimated using LINKNE for each model based on 1000 SNP loci and a sample size of 50 except where the change in N_e generated a population of less than 50 individuals (in which case all individuals were sampled). Change in N_e was detected when confidence intervals between estimates from the most distant past (14.3 generations) and those from the most recent past (1.01 generations) did not overlap.

Demographic Model	Initial N_e	Change in N_e Detected
Constant	100	No
	250	No
	500	No
	1000	No
Decline to 75%	100	Yes
	250	Yes
	500	No
	1000	No
Decline to 50%	100	Yes
	250	Yes
	500	Yes
	1000	No
Decline to 25%	100	Yes
	250	Yes
	500	Yes
	1000	Yes
Expansion to 2x	100	Yes
	250	Yes
	500	Yes
	1000	No
Expansion to 5x	100	Yes
	250	Yes
	500	Yes
	1000	Yes

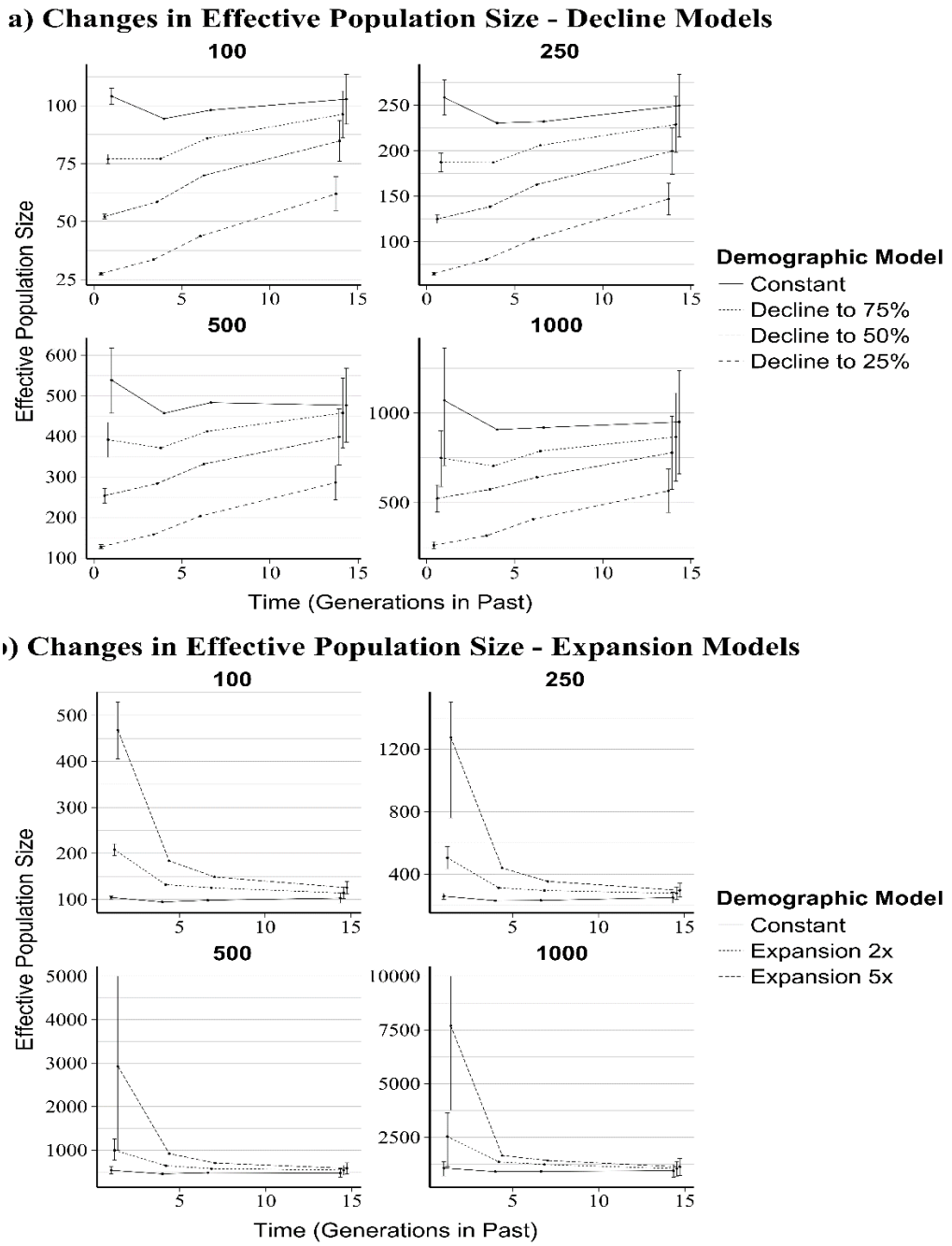


Figure 3.1 Estimates of N_e over time in the past for various demographic models, calculated with LINKNE. Each panel represents a different starting effective population size (N_e). Confidence intervals are shown only for estimates from the most recent and most distant past. Time points within a plot are adjusted horizontally so that confidence intervals could be distinguished. (a) Trend lines for constant and decline models. (b) Trend lines for constant and expansion models. Upper confidence limits for estimates of N_e for the expansion to 5x model and for starting N_e of 250, 500, and 1000 were truncated for clarity and are marked with an arrow to indicate that the interval extends beyond the limits of the plot

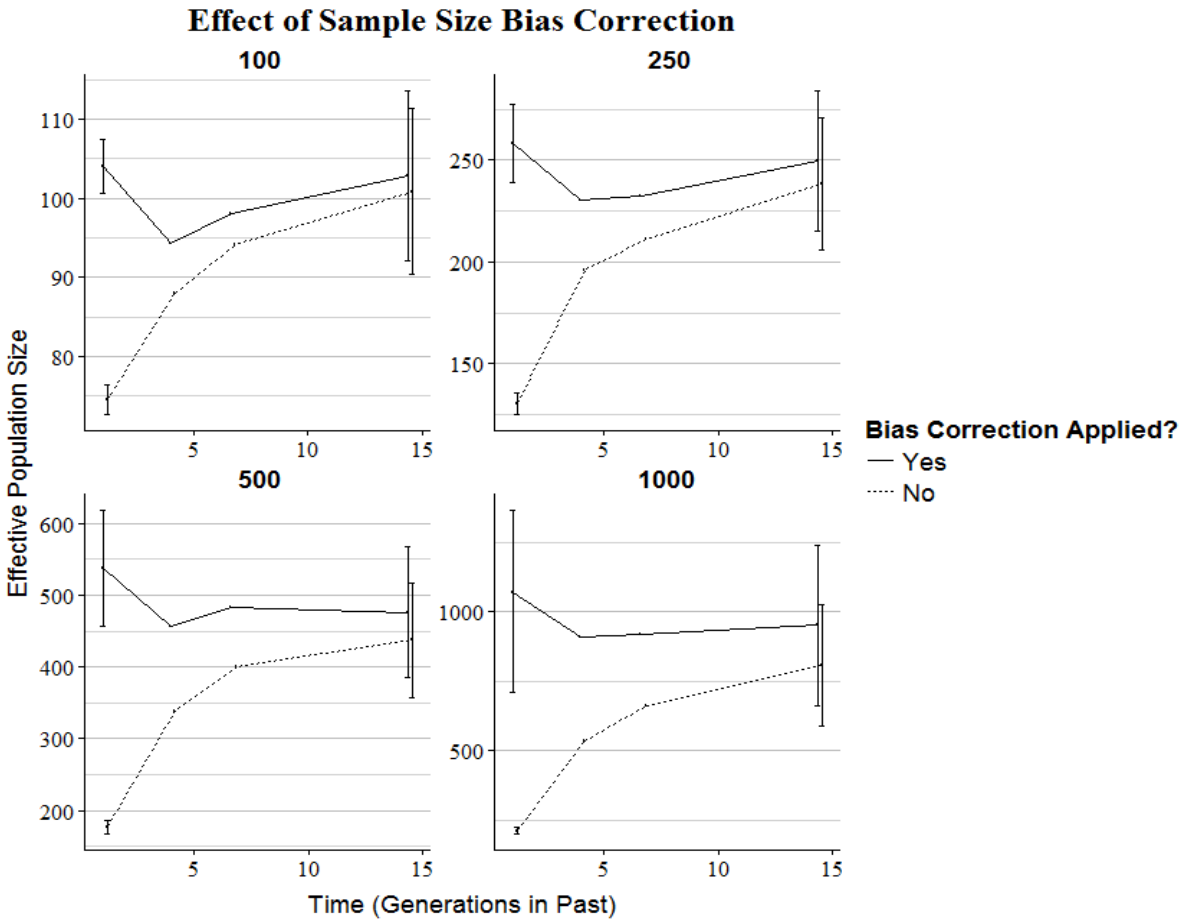


Figure 3.2 Effect of sample-size bias correction. Calculations used the sample-size (*S*) bias correction proposed by Waples (2006) on estimates of N_e over time. Estimates were produced with LINKNE. Solid lines represent estimates of N_e, with bias correction applied, for the constant population size model. Dashed lined represent estimates where r^2_{sample} was measured as 1/*S*.

at 1.01 generations in the past were fairly accurate as bias for models of decline to 25%, 50%, and 75% (averaged across simulations for all starting values of N_e) were 4.06%, 2.15%, and 1.35%, respectively. Estimates of N_e for the most distant time in the past (14.3 generations) were downwardly biased for all decline models; bias for declines of initial N_e to 25%, 50%, and 75% were -41.5%, -19.6%, and -8.71%, respectively.

Expansions in N_e (Figure 3.1b) were detected in all but one model (initial N_e of 1000 and a 2x expansion); confidence intervals between the most recent (1.01 generations) and most distant (14.3 generations) times in the past overlapped slightly. Bias for estimates of N_e at 1.01 generations in the past, averaged across starting values of N_e, was positive and less than 10% for expansions of 2x and 5x (7.88% and 1.95%, respectively); bias varied considerably over each time period for different values of N_e (2x: -0.23% to 27.08%; 5x: -41.5% to 53.77%). Estimates of N_e in the past were influenced less by expansions in population size than by declines.

Evaluation of Sample-Size Bias Correction

The sample size bias proposed by Waples (2006) influenced estimates of N_e in the recent past to a greater extent than estimates in the more distant past (Figure 3.2). When the bias correction was not applied, a downward bias was present for estimates at all points in time and was larger for more recent time periods, with an average bias of -7.93% (14.3 generations), -18.7% (6.65 generations), -26.9% (3.99 generations), and -50.9% (1.01 generations). There also was an effect of N_e on bias, as downward bias increased with increasing N_e. Overall, failure to apply the bias correction resulted in a significant

downward trend, falsely indicating that the model of constant size had experienced a recent decline in N_e.

Allele Frequency Cutoff

The cutoff value for excluding rare alleles had the most influence on estimates of N_e from the most distant past (Figure 3.3). For estimates of N_e in the recent past (1.01 generations), mean values of N_e ranged from 252.7 to 274.4 (range = 26.74); for estimates of N_e in the most distant past (14.3 generations), values ranged from 232.2 to 276.5 (range = 44.28). For all allele-frequency cutoff values evaluated, estimates of N_e in the recent past were upwardly biased, while the direction of bias for estimates in the more distant past depended on the level of cutoff chosen. No cutoff value was the least biased for all time points, although a cutoff value of 0.05 appeared to be the best compromise, as it resulted in the least bias, on average, across all time points (Figure 3.3).

Effect of Time between Demographic Change and Sampling

The number of generations between demographic change and sampling had a large effect on resulting estimates of N_e (Figure 3.4). For all demographic change models, estimates of N_e derived from unlinked and moderately linked loci ($c > 0.15$ M) equilibrated to the correct N_e within five generations; whereas estimates from tightly linked loci ($c \leq 0.15$ M) approached the new N_e more slowly. Both population expansions and declines could be detected up to 20 generations in the past. Estimates from the distant past (14.3 generations) tended to equilibrate more slowly for demographic expansions than for declines.

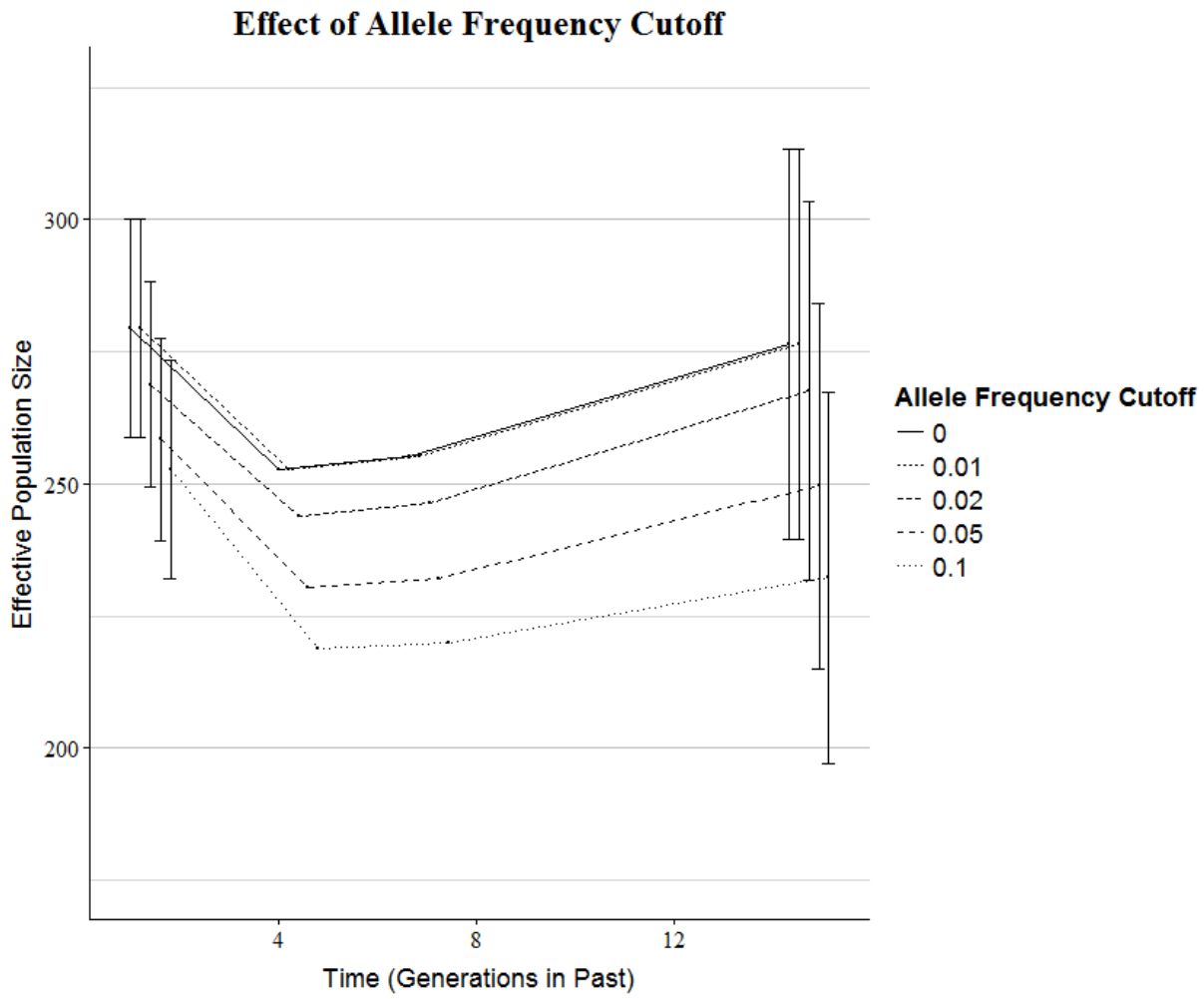


Figure 3.3 Effect of excluding rare alleles at various thresholds. Estimates of N_e were produced using LINKNE and various rare-allele exclusion thresholds (0.10, 0.05, 0.02, 0.01, and 0), using the constant population model with N = 250.

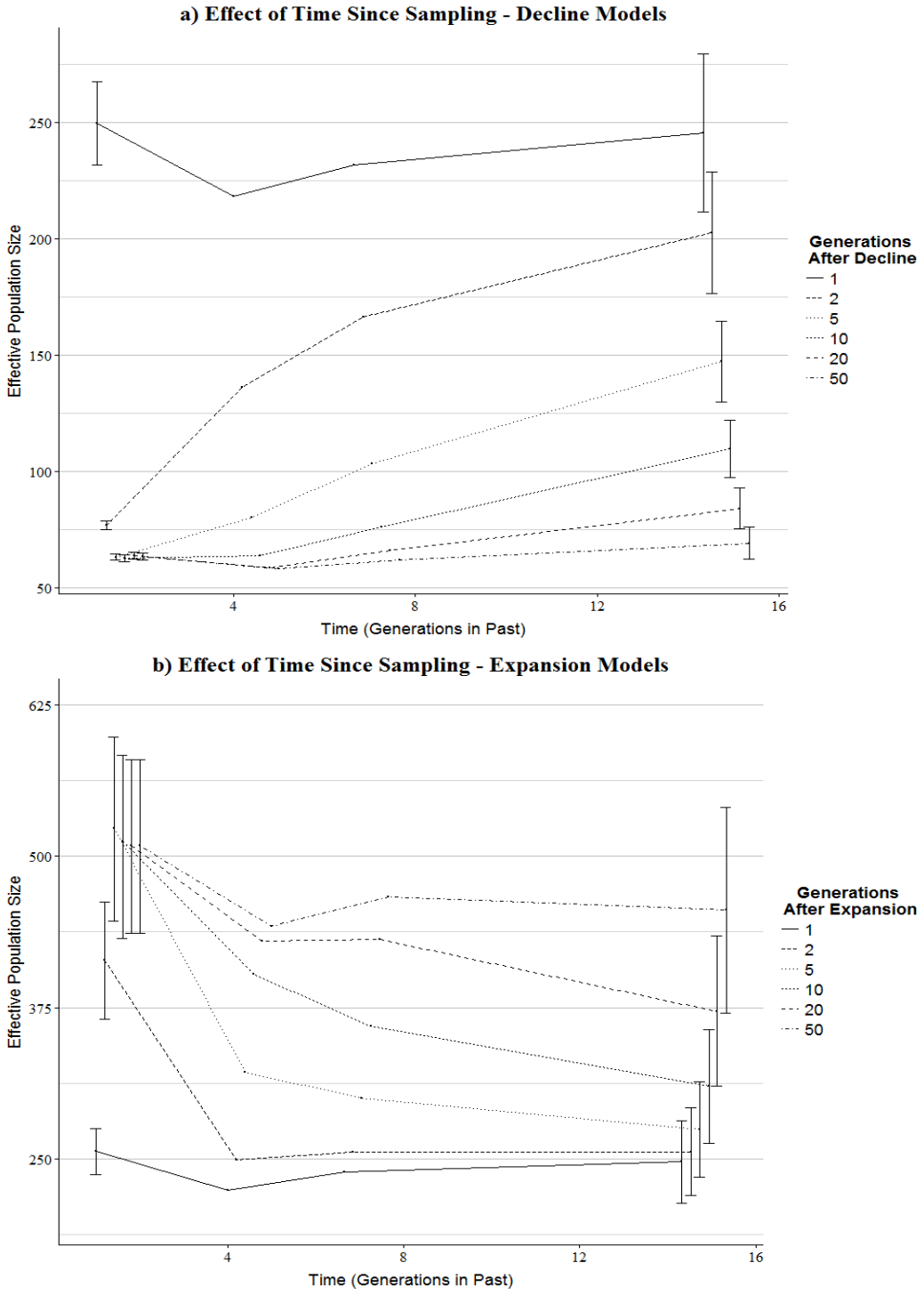


Figure 3.4 Effect of length of time between demographic change and sampling. (a) Results for decline to 25% of initial N_e of 250 when sampling was conducted 1, 2, 5, 10, 20, and 50 generations after a decline. (b) Results for an expansion to 2x of initial N_e of 250 when sampling was conducted 1, 2, 5, 10, 20, and 50 generations post-expansion.

Comparison to the LDNe Method

For all values of N_e under the constant model, estimates of current N_e based on NEESTIMATOR2 were biased downward by more than 20% (Figure 3.5). Bias for initial N_e of 100, 250, 500, and 1000 was -25.4%, -23.0%, -20.9%, and -21.2%, respectively. Estimates generated using LINKNE had a small upward bias of 3.54%, 3.15%, 7.51%, and 6.96% for initial N_e of 100, 250, 500, and 1000, respectively.

Empirical Data

The trend line for the sample from West Matagorda Bay (Figure 3.6a, dashed line) was suggestive of a recent decrease in N_e, consistent with the presence of hatchery-raised individuals in the sample. Separating the sample into hatchery-raised and wild fish revealed that estimates of N_e over time for wild fish were large and featured no observable trend (Figure 3.6a, grey ribbon); estimates from hatchery-raised individuals alone (Figure 3.6b) were consistent with the expected bottleneck (based on Gold *et al.*, 2008) of progeny from the parental brood stock. Trend lines for both the mixed sample and the hatchery-raised individuals were consistent with results of simulations (see Figure 3.1a, decline to 25%); estimates of N_e for the more distant past appeared lower than expected and the slope of the trend line less steep due to the recent effect of increased genetic drift on tightly linked loci.

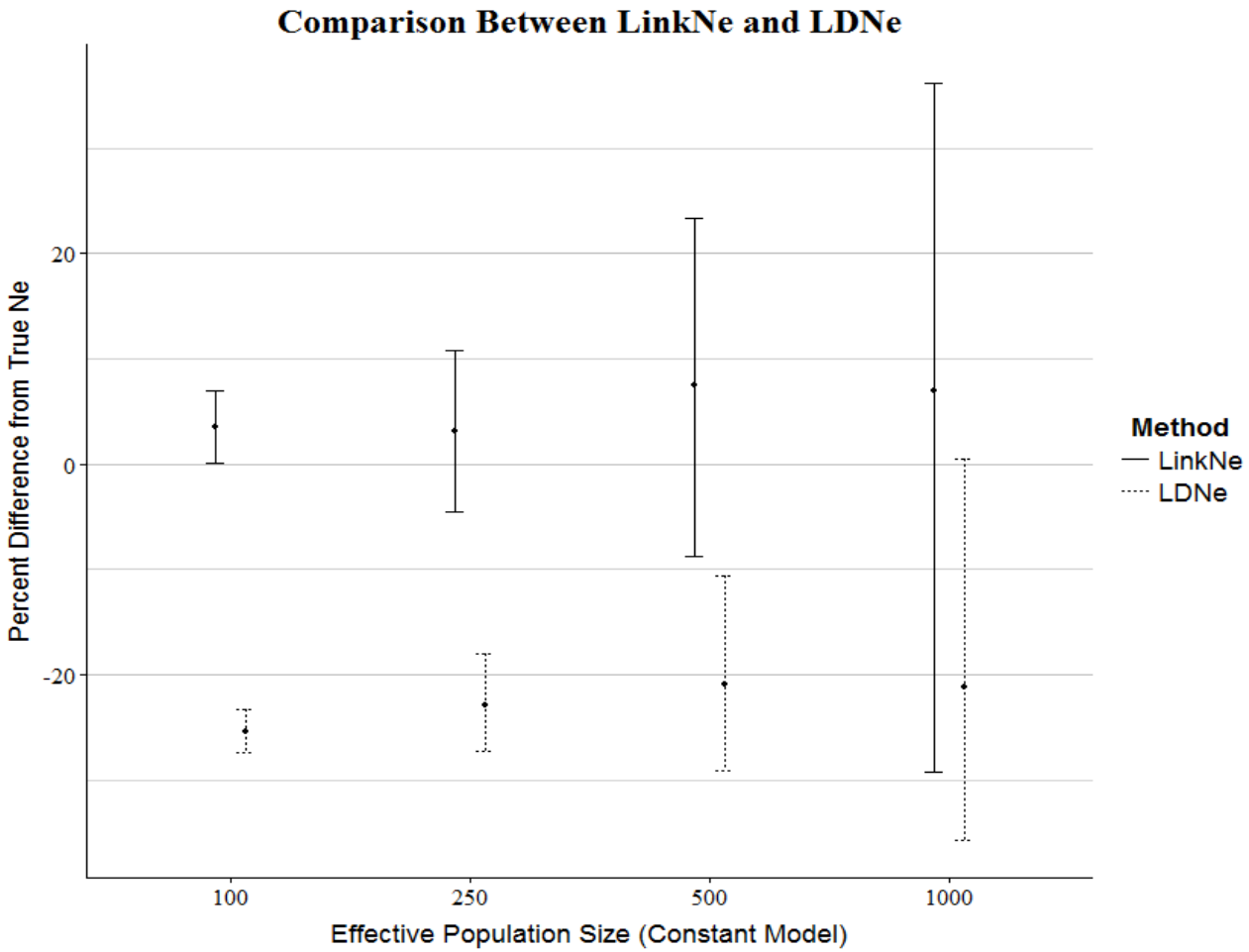


Figure 3.5 Comparison of bias in measures of N_e, using LINKNE and the LDNE method (as implemented in NEESTIMATOR2). Bias was measured as the difference between the estimated and true N_e and is expressed a percentage of the true N_e. Estimates of N_e based on linkage disequilibrium can be biased downward when linked loci are assumed to be unlinked.

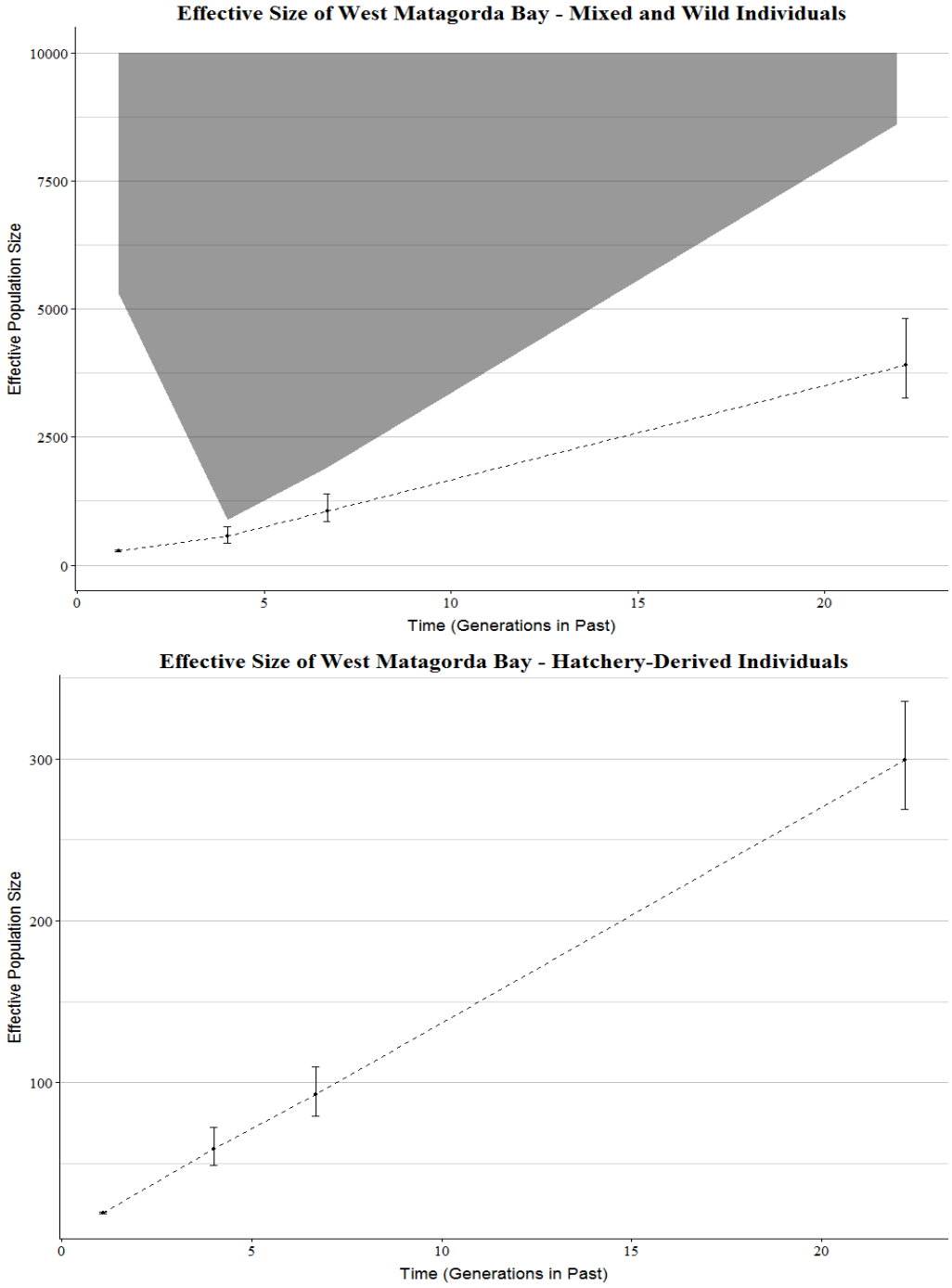


Figure 3.6 Results of analysis of a sample of red drum juveniles from Matagorda Bay, TX. (a) Trend line (dashed) for N_e produced using LINKNE and all sampled individuals and confidence interval (CI) of N_e when F_1 hatchery individuals are removed (shaded area). Note that when 'wild' individuals are assessed, only the lower bounds of the CI are estimable from the data; for clarity, the CI is truncated at 10,000. (b) Trend line for F_1 hatchery-raised individuals, indicative of a large decline in N_e in the parental generation, which consisted of hatchery brood stock.

DISCUSSION

Simulation

Precision and Bias

The ability of this or any approach to identify changes in N_e over time is largely dependent on precision and potential bias. If estimates of N_e at different times in the past are systematically biased, inferences regarding demographic trends will be compromised. Results of simulations revealed <10% bias in estimates of N_e for populations of constant size over the time period (~1-15 generations in the past) assessed. However, the magnitude and direction of the bias depended on both the time to which an estimate referred and the true N_e. This suggests that while the precision provided by the number of simulated loci (1000) was such that confidence intervals for estimates of N_e across time tended to overlap for the constant population at all initial effective sizes, increasing the number of loci could produce estimates so precise that confidence intervals would not overlap, even for populations of constant size. However, because bias for all estimates was small (<10%), it would be unlikely that such a situation would be confused for a large change in N_e.

Further study is needed to evaluate the source of bias at different periods in time. For example, it is not clear why estimates from intermediate time points tend to be more biased and in a downward direction. It should be noted that in addition to the sample-size bias correction, a simulation-based bias correction for the drift component of r^2 was proposed by Waples (2006) for unlinked loci. An applied correction for linked loci might eliminate some of the bias, but the correction factor would be challenging to

implement because a correction factor would have to be calculated for all values of *c* across the spectrum of possible linkage values. While Waples (2006) found little bias in N_e due to drift for unlinked loci when initial N_e was greater than 100, the smallest initial N_e evaluated in our study, it is unclear whether this also is true for linked loci.

Our findings regarding precision of N_e estimates over time are in agreement with Hill (1981) who showed that the coefficient of variation of N_e decreases as the recombination rate decreases and the number of pairwise locus comparisons increases. This means, given an equal number of pairwise comparisons, that estimates of N_e in the past should be more precise than recent estimates (Hill, 1981; Hayes *et al.*, 2003). However, the vast majority of locus pairs in a genome are unlinked, so the large number of pairwise comparisons available should yield recent estimates with a high level of precision. Consistent with this, intermediate time periods (corresponding to intermediate values of *c*) had the lowest level of precision, most likely as a consequence of having the fewest number of pairwise comparisons.

Detection of Changes in N_e

Results of simulations demonstrated that for ideal populations, recent changes in N_e can be reliably detected by comparing estimates of N_e based on LD from pairs of linked and unlinked loci. In our simulations, trend lines for the constant population at all initial effective sizes never indicated a change in N_e, although trend lines for models with a change in N_e in some models indicated stability. This has important implications for interpretation of results when using the linkage approach as it indicates that although detected changes in N_e are robust, results indicating constant size need to be carefully

scrutinized. Our simulations revealed that changes in N_e are more readily detected when N_e is small, largely due to increased precision of LD-based estimators at smaller N_e. In fact, even relatively small changes in N_e (declines to 75% of the original value) were detected provided that the initial N_e was 250 or less. The linkage approach was less effective in populations with larger initial N_e as only changes in N_e of relatively large magnitude could be detected. However, increasing the sample size, which was fixed at 50 individuals for all simulations, should improve resolution to detect changes for populations of larger initial N_e.

Estimates of N_e were fairly accurate for more recent time ($t \approx 1$ generation) in the past. This is because sampling was conducted five generations after the change occurred and unlinked (or moderately linked [$c > 0.15$ M]) loci are expected to equilibrate to the new steady-state level of LD in three to four generations (Sved, 1971; Waples, 2006). Estimates from the more distant past (>3.33 generations) reflected N_e of the population prior to the change in N_e; however, these estimates tended to be influenced by more recent N_e. This effect was particularly pronounced for decline models, where estimates reflecting prior generations showed a considerable downward bias, causing trend lines to be less steep than expected. In addition, the bias was exaggerated for declines of large magnitude. Estimates of N_e in the past during expansion models were less influenced by more recent N_e, and it is likely that the different effects on estimates of N_e in the past observed in decline and expansion models relate to the relative contribution of drift and recombination to steady-state levels of LD. In the case of a decline, LD accumulates between loci at every linkage interval relatively quickly due to the increased importance

of drift. Alternatively, in an expanding population drift becomes less important as time is required for recombination to dissolve LD between linked loci. In practice, this suggests that the true magnitude of a decline in N_e would be difficult to detect with certainty because past estimates would be influenced by effects of drift in more recent generations; estimates of past N_e following a population expansion, however, may provide a more reliable estimate of the magnitude of the change in N_e.

A critical component of the linkage approach is establishment of a relationship between recombination rate and time. While an approximate relationship was suggested by Hayes *et al.*, (2003), it was derived under the limiting assumptions that c is small and that N_e changes linearly with respect to time. Despite the fact that these assumptions are clearly violated, trend lines from our simulations agreed reasonably well with known timing of changes in N_e, particularly for expansion models. The results were less concordant for decline models, as trend lines suggested more gradual declines than expected. This is likely due to effects of increased genetic drift following a decline. Organizing locus pairs into bins and using a mean value for c, while necessary for achieving acceptable levels of precision, is one source of discordance between theoretical and observed results. Depending on the size of the bin and the degree of linkage, estimates of LD at locus pairs in genomic regions reflecting N_e across multiple generations are collapsed into a single estimate which may obscure fine-scale trends.

The simulations evaluated consisted of ideal populations with non-overlapping generations, even sex ratios, and binomially distributed reproductive success, such that $N \approx N_e$. More rigorous investigation is necessary to evaluate effects on estimates made

when these assumptions are violated. Effects of skewed sex ratio and increased variance in reproductive success on estimates of contemporary N_e generated with the LD method have been investigated to some extent by Waples (2006), with the conclusion that the assumptions are fairly robust to the influence of these effects, i.e., an ideal population with a given N_e is a reasonable proxy for non-ideal populations with the same N_e (Waples 2006). However, the biological characteristics of the species tend to determine the N_e/N ratio (Portnoy *et al.,* 2009), and it is likely that changes in census size (N) influence estimates of N_e differently. Therefore, while the linkage method can robustly detect changes in N_e, care must be taken when interpreting the results in terms of changes in census size. Additional study will be necessary to understand the influence of other factors that shape patterns of genome-wide LD in natural populations on estimates of past N_e; these factors include selection, migration, admixture, and complicated demographic models.

Effect of Sample Size Bias Correction

Our simulations demonstrate the importance of sample-size (S) bias correction for accurately assessing changes in N_e. England *et al.,* (2006) and Waples (2006) demonstrated that estimates of N_e can be downwardly biased when S is small relative to the true N_e. Our simulations showed that this bias is more pronounced for estimates of N_e in the more recent past. When bias correction was not applied, the linkage method produced trend lines characteristic of a decline in N_e, even for the constant populations that had not experienced a decline. This is an important consideration, and little attention has been given to the effects of S on estimates of N_e in studies applying similar

methods. It is important to note that the effect of *S* may be dependent on the way in which r^2 is estimated, with estimators where marker phase is known requiring a smaller correction factor (Corbin *et al.,* 2012).

Allele Frequency Cutoff

The effect of modifying the cutoff value for excluding rare alleles varied depending on the time in the past to which estimates applied and there was no single, optimal allele-frequency cutoff. In general, a cutoff at an allele frequency of 0.05 produced estimates of N_e across the range of time points that were closest to the true N_e. In addition, results for estimates based on unlinked loci were consistent with the findings of Waples and Do (2008) which indicated that larger cutoff values minimized upward bias caused by occurrence of rare alleles. Further, our results paralleled that of a previous study (Corbin *et al.,* 2012), where effects of modifying rare allele cutoffs for estimates of past N_e, using phase-known data, was explored. It was concluded (*Ibid*) that a cutoff value between 0.05 and 0.1 produced the most accurate estimates. Applying a separate cutoff to locus pairs in different bins may produce more accurate estimates across all time points, if increased cutoff values were used for estimates further back in time.

Effect of Time since Sampling

Several insights were gained by modifying time of change in N_e relative to sampling. First, based on evaluating overlap of confidence intervals between past and present estimates, the linkage approach was able to detect both expansions and declines in N_e at least 20 generations in the past. In theory, it is possible to obtain estimates of N_e in the

much more distant past (and to thus detect older demographic changes) if LD can be measured between very tightly linked loci (< 0.01 M). However, simulations by Corbin *et al.*, (2012) suggest that estimating long-term trends can be problematic, in part because the effect of mutation is important over long periods of time. Second, analysis of trend lines for decline and expansion models reinforced the idea that past estimates of N_e are influenced more by declines than expansion, as past estimates of N_e rapidly approach the new steady-state level of LD after a decline but approach the new level more slowly following an expansion. When the change in N_e occurred 50 generations in the past, neither declines nor expansions could be statistically differentiated from stasis. In the case of declines, the mean estimate of N_e was the same between the most recent generation and the generation furthest in the past. For expansions, the mean estimate of N_e was larger in the most recent generation than the generation furthest in the past; however, precision was limiting and confidence intervals overlapped. This further suggests that genomic patterns of LD indicating an expansion in N_e persist for longer, enabling expansions in the more distant past to be detected.

Comparison with LDNE method

Results from simulations indicated that the assumption that all loci in a genome-wide data set are unlinked can downwardly bias estimates of contemporary N_e by as much as 25%. In the absence of marker linkage or genomic position data, it is unclear what should be the best strategy for avoiding this bias. One approach is to remove estimates from locus pairs with excessively high LD as they possibly are influenced by physical linkage (Gruenthal *et al.*, 2014); in practice, however, the decision to remove such loci is

fairly arbitrary. Regardless, in the absence of known linkage relationships, acknowledging that estimates of N_e from the LDNE method likely underestimate the true value is a conservative approach; the fact that the bias is downward is favorable from a biological risk-assessment standpoint because overestimating N_e likely will have more dire consequences for imperiled species than underestimating N_e.

Empirical Data

A decrease in N_e in the sample of juvenile red drum from Matagorda Bay in 2008 was detected using the linkage approach. Presumably the decline in N_e was due to the presence of an inordinately large proportion of hatchery-raised juveniles in the sample. The effect, as expected, was temporary as the current N_e of a second sample from the same locality, taken in 2015, was considerably larger than the estimate of current N_e in the wild fish in the 2008 sample (unpublished data). This highlights that interpretation of trends in N_e based on LD should be made with caution. If the trend line for the mixed sample had been generated with no knowledge about the constituents of the population, one might have hypothesized erroneously that the population of red drum in Matagorda Bay had experienced a recent, large decline in N_e possibly caused by a decline in census population size rather than an unequal contribution of progeny from a limited number of breeders in the parental generation. Additionally, despite the rapidity of the decrease in N_e, the trend line suggested a more gradual decline that extended into the distant past. This likely occurred for several reasons, including uncertainty in estimating

recombination rates from the mapping cross and the necessity of binning loci over large genomic distances.

Another important consideration when evaluating potential changes in N_e using large, empirical datasets is that parametric confidence intervals (calculated based on a chi-square approximation [Hill, 1981]) may be too narrow when many loci are utilized (Waples *et al.*, 2015). This is because, as the number of utilized loci increases, there are more correlations among r^2 values for locus pairs that share common loci, and this increasingly violates the assumption of independence of comparisons implicit in the parametric model (Waples, 2006; Waples and Do, 2010). As a result, parametric confidence intervals do not adequately convey the uncertainty in r^2, and the standard jackknife procedure for correcting confidence intervals (Waples and Do, 2008) will not alleviate the problem when a large number of loci are used (Do *et al.*, 2014). Because the linkage method presented here tends to separate pairwise comparisons from the same locus into different bins, the effect is likely relatively less pronounced as compared to a single estimate of N_e using all loci when linkage data is unavailable. However, when comparing confidence intervals of N_e across different points in time it is important to consider the possibility of overly precise and inaccurate estimates. Further study will be needed to quantify the extent to which this problem affects genome-scale datasets. Regardless, considering that bias appears to be relatively low, overly tight confidence intervals are unlikely to result in false detection of large changes in N_e.

Conclusions

We have shown that when linkage or genomic position data are available, the LD approach of estimating N_e from unphased genetic markers (Waples, 2006; Waples and Do, 2008) can be extended to estimate N_e in the recent past and, importantly, to detect recent changes in effective population size (N_e). Results of simulations suggested that even with a moderate number of loci, relatively small changes in N_e (25%) could be detected provided that initial N_e was not large. Further, we explored the effects that sample-size bias correction, rare allele cutoff, and time since the change occurred have on estimates of N_e across points in time and quantified the bias in N_e associated with the assumption that all SNPs in a genome-wide dataset are unlinked. Finally, we demonstrated the utility of the linkage method for detecting recent changes in N_e on an empirical data set. Overall, results of the analysis of both simulated and empirical data suggest that this approach will be useful for genetic monitoring, particularly when prior genetic samples are not available. This strategy should become increasingly available to species of conservation concern as genotyping-by-sequencing techniques are widely adopted and as genome sequences and linkage maps become more available.

CHAPTER IV

A POPULATION GENOMIC ASSESSMENT OF RED DRUM IN US WATERS AND CONCLUSION

INTRODUCTION

Advances in next-generation sequencing (NGS) technology have enabled cost-effective screening of thousands of genetic markers for nearly any species (Davey *et al.,* 2011). One of the principal benefits of dense, genome-wide sampling of genetic markers in wild populations is the ability to quantify the relative effects of micro-evolutionary phenomena that impact the genome as a whole (e.g., genetic drift and migration) and phenomena that have locus-specific impacts (e.g., natural selection) (Luikart *et al.,* 2003). Population-genetic theory and empirical studies indicate that genomic regions under the influence of selection should show elevated divergence with respect to the rest of the genome and that only a small proportion of the genome may show signatures of selection (Lewontin and Krakauer, 1973; Wu, 2001; Nosil *et al.,* 2008). Genome scans, using hundreds or thousands of genetic markers, can sample enough of the genome to detect and localize regions that contain adaptive genetic variation (Luikart *et al.,* 2003, Allendorf *et al.,* 2010).

The ability to efficiently recover marker position information, either in the form of a genome sequence or a genetic map (Baird *et al.,* 2008), offers a major advantage towards elucidating these processes. Linkage mapping is a cost-effective alternative to genome sequencing and, used in conjunction with population-genetic data, can provide

each locus a genomic position and allow for the visualization of population-genetic statistics as continuously distributed variables across a genome (Hohenlohe *et al.,* 2010a, 2010b). This information in turn can be used to identify regions of the genome that are potentially under selection, can minimize the proportion of false positive genomic outliers, and enable use of comparative genomic approaches to identify potential candidate genes in proximity to genetic markers identified as potentially being under the influence of selection (Akey *et al.,* 2002).

The identification of genomic regions under the selection has important implications for exploited species (Bradbury *et al.,* 2013). In the context of management and conservation, adaptive genetic variation is important because it represents the component of the genome that allows individuals to survive and reproduce in local environmental conditions. Characterizing genomic regions experiencing selective pressures in response to the environment is therefore important for establishing appropriate units of conservation (Waples, 1995; Fraser and Bernatchez, 2001). In addition, because adaptive variation to some extent represents the evolutionary potential of a population (Waples, 1995), identification of adaptive variation is important for monitoring of a population's potential for response to environmental change or exploitation (Schwartz *et al.,* 2007; Allendorf *et al.,* 2008). Moreover, adaptive variation is useful in the context of population-structure analysis because inclusion of loci under divergent selection can increase power to discriminate between populations that experience high levels of gene flow, which tends to homogenize neutral variation, or

that have diverged too recently for drift to have appreciably influenced neutral allele frequencies (Nielsen *et al.,* 2009).

Red drum (*Sciaenops ocellatus*) is an estuarine-dependent, marine fish species in the U.S. Gulf of Mexico (hereafter Gulf) and U.S. South Atlantic (hereafter Atlantic). The species supports one of the largest recreational fisheries in the U.S. (NMFS, 2015b). Assessment of genetic structure of red drum populations in the Gulf and Atlantic began with assays of allozyme polymorphisms nearly 30 years ago (Ramsey and Wakeman, 1987; Bohlmeyer and Gold, 1991), with subsequent studies utilizing improvements in genetic marker technology and population genetics analysis, including mitochondrial RFLPs (Gold and Richardson, 1991; Gold *et al.,* 1993, 1999), mitochondrial sequencing (Seyoum *et al.,* 2000), and microsatellites (Chapman *et al.,* 2002; Gold and Turner, 2002). The consensus regarding population structure among prior studies is the existence of weak, but significant genetic differentiation between Gulf and Atlantic populations of red drum (Gold and Richardson, 1991; Gold *et al.,* 1993, 1999; Seyoum *et al.,* 2000; Chapman *et al.,* 2002), and an isolation-by-distance pattern of differentiation among red drum in the Gulf (Gold *et al.,* 1999, Gold and Turner 2002). Gold *et al.,* (2001) reviewed the status of population genetics in red drum and hypothesized that because red drum rely on discrete estuarine habitat for successful reproduction and recruitment, populations in the Gulf could be described by a modified, one-dimensional, stepping-stone model where gene exchange occurs primarily between adjacent estuaries. It could be further hypothesized that given the species' dependence

on estuarine habitat, which tend to be environmentally heterogeneous (Schulte, 2007), the potential for local adaptive differences could be large.

The goal of this study was to conduct a more comprehensive population genomics assessment of red drum in U.S. waters, using high-resolution sequencing technologies, with the following specific objectives: i) to describe patterns of both neutral and putatively adaptive genetic variation in the species across the sampled range, and ii) to utilize a genetic linkage map, combined with population genomics data, to examine genomic patterns of local adaptation in the species.

MATERIALS AND METHODS

Reduced Representation Reference Genome Construction

Prior to analysis of genetic data, a reduced representation, reference genome was constructed for red drum. Twenty red drum individuals from across the species' natural range, including parents from two mapping crosses, were used to produce a double-digest restriction-site associated DNA (ddRAD) library, following standard procedures (Peterson *et al.,* 2012). The library was sequenced on an Illumina MiSeq DNA sequencer producing 300 bp, paired-end reads. The raw sequencing reads were demultiplexed, using the program process_radtags from the *Stacks* package (Catchen *et al.,* 2011), and a *de novo* reference genome was assembled with the *dDocent* pipeline (Puritz *et al.*, 2013). Because sampled RAD contigs had a mean size of 300 bp, the entire sequence of each RAD contig was recovered during reference assembly. A preliminary annotation of the reference genome with the BLAST algorithm revealed the

presence of multi-copy nuclear ribosomal RNA genes. To avoid problems with read mapping cause by multi-copy genes, a custom script was used to remove the rRNA contigs, as well as all contigs with a total length of less than 150 bp. All further RAD sequencing analysis utilized this reference genome for read mapping and SNP calling.

Linkage Map

Sample Collection and Genotyping

Tissue samples from the two, full-sibling mapping crosses used for generating a microsatellite linkage map (Portnoy *et al.,* 2010, Hollenbeck *et al.,* 2015) were extracted using Mag-Bind Tissue DNA kits (Omega Bio-Tek). RAD libraries were constructed following procedures outlined in Chapter 3 and sequenced on two lanes of an Illumina HiSeq 2000 DNA sequencer. Raw sequencing reads were demultiplexed, using the program process_radtags, to produce a file containing the raw reads for each individual. Read mapping and SNP calling were conducted for each family separately, using the *dDocent* pipeline and the reference genome described previously. Raw SNP genotypes were stringently filtered based on numerous criteria, using the VCFtools package (Danecek *et al.,* 2011). First, individual genotypes called from less than ten reads were removed, followed by all loci with a mean Phred quality score of <20. An iterative filtering process was then used to maximize the number of individuals and loci in the final dataset. Loci with >50% missing data were excluded, followed by individuals with a mean depth of less than five reads and >95% missing data. Next, loci and individuals with >25% missing data were then excluded, and loci with a minor allele frequency <

0.05 were removed. Next, the bash script dDocent_filters was used to remove loci based on numerous criteria, including mean site depth, the ratio of quality to depth, strand representation, allelic balance in heterozygous individuals, and proper read pairing. Complex polymorphisms were then decomposed to individual SNP or indel loci, using the vcfallelicprimitives command in the vcflib package. Next, loci were collapsed into haplotypes within each RAD contig, using the program rad_haplotyper. The program removed loci that were successfully haplotyped in <75% of individuals and kept indel loci when the indel was the only polymorphism on the contig; other indels were excluded from the analysis to avoid complications in haplotyping. The resulting file for each family contained one diploid genotype per individual for each remaining RAD contig. A custom script was used to convert the rad_haplotyper output files for both families to JOINMAP 4.1 (van Ooijen, 2012) format.

Linkage Mapping

RAD-based genotype data was combined with microsatellite genotype data obtained previously (Portnoy *et al.*, 2010, 2012; Hollenbeck *et al.*, 2015), and the resulting data file imported into JOINMAP. RAD loci were then added to previously defined linkage groups (Portnoy *et al.*, 2010, 2012; Hollenbeck *et al.*, 2015). To reduce the chance of incorrectly adding loci to existing groups, an initially conservative LOD score of 9.0 was applied, followed by two more rounds of grouping at an LOD of 6.0 and 3.0, respectively. After assigning loci to linkage groups, groups of loci that exhibited parallel genotype distributions (an observed recombination rate = 0) were identified and only a

single representative locus was retained for initial ordering; the remaining loci were reserved for later addition to the map. Tests for segregation distortion were carried out using chi-square goodness of fit tests, using JOINMAP. Family-specific maps were generated by applying the multipoint maximum likelihood algorithm for outbred crosses, as implemented in JOINMAP. Marker order for shared loci was compared between families, and incongruences were corrected by identifying and removing problematic loci, which most often displayed segregation distortion or contained probable genotyping errors. Loci excluded initially because of lack of recombination with another locus were then added to family-specific maps by placing them into the same map position as the representative mapped locus. Family-specific maps were then combined into a consensus sequence with the program MERGEMAP (Wu *et al.*, 2011), using equal weights for both families.

Population Genomics

A total of 563 juvenile (0-3 year old) red drum were sampled between 2008 and 2015 from 11 localities in the Gulf and Atlantic. Sample localities (Figure 4.1) included three localities in Texas: Lower Laguna Madre (LLM), Matagorda Bay (MAT), and Sabine Lake (SAB); Mississippi (MIS); four localities in Florida (Apalachicola, APA; Cedar Key, CEK: Charlotte Harbor, CHA; and Indian River, IND); two localities in Georgia (Hampton River, HAR; and Wassaw Sound, WAS); and one locality in South Carolina

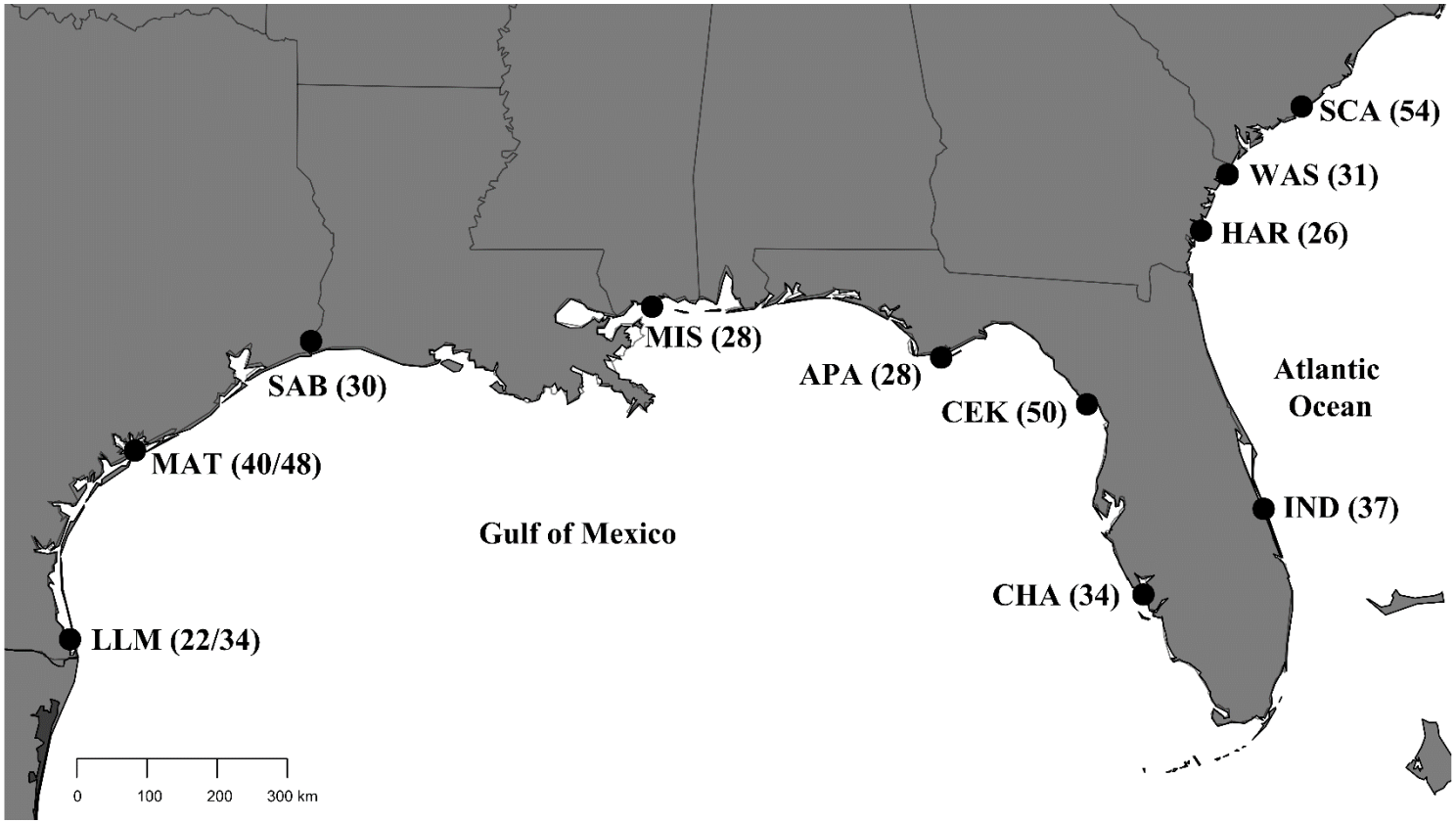


Figure 4.1 A map of 11 sampling localities for red drum. Numbers in parentheses represent the final, filtered sample size for each locality. The two sample sizes for LLM and MAT represent the size of each temporal sample from both localities taken in the spring of 2008 (first value) and the winter of 2014/2015 (second value).

(SCA). Because all samples consisted of juvenile fish, genetic differences observed between samples could arise in part by random recruitment variability in a particular year class. To gauge the potential magnitude of between year variation in allele frequencies, two localities in Texas, LLM and MAT, were sampled in both 2008 and 2014/2015 and compared to assess temporal stability of genomic variation.

ddRAD Library Preparation and Bioinformatics Filtering

RAD libraries were prepared as outlined in Chapter 3 and sequenced on five lanes of an Illumina HiSeq 2000 DNA sequencer. Demultiplexing, read mapping, and SNP calling were performed as above, and SNP loci were filtered, using VCFtools. Individual genotypes called with less than ten reads were excluded, as were all loci with a mean Phred score of <20. Next, loci with >50% missing data were then excluded, followed by all individuals with a mean locus depth of less than ten and >75% missing data. Loci with >25% missing data were then excluded, followed by individuals with >40% missing data. Loci with a minor allele frequency of <0.05 were then excluded. Next, genotypes of five individuals, intentionally duplicated in different RAD libraries, were compared for genotype discordance. Loci which exhibited genotype differences across more than one set of replicates were discarded, and the individual from each set of duplicates that contained the most missing genotypes was subsequently removed from the data set. Following these steps, suspected first-generation, hatchery-derived individuals in the samples from Texas waters were removed from the dataset. Hatchery-derived individuals from the Texas stock-enhancement program were identified based on

data from a previous study (Carson *et al.*, 2014) or from high genetic relatedness (r > 0.35; Yang *et al.*, 2010) to at least one current hatchery brood fish; the latter had been genotyped with the same set of SNP loci for an ongoing study. Because hatchery-derived juveniles could only be identified directly in samples from Texas waters, highly related individuals from other samples, which potentially could reflect hatchery origin, also were identified using the method of Yang *et al.*, (2010), as implemented in VCFtools. One individual from each pair with a relatedness coefficient (r) > 0.2 and sampled from the same locality was removed from the dataset. Next, the *dDocent* filters script was used to remove loci based on mean site depth, the ratio of quality to depth, strand representation, allelic balance in heterozygous individuals, and proper read pairing. A maximum mean depth cutoff of 300 reads was applied in the script. At this point, complex polymorphisms were decomposed to individual SNP or indel loci, using the vcfallelicprimitives command in the vcflib package. Next, loci were tested for conformance to Hardy-Weinberg equilibrium within each locality, and loci failing (P < 0.001) in more than two of the 13 samples were excluded. Individuals with >25% missing data were then excluded, followed by loci >15% missing data in any single locality and more than five percent missing data across the entire data set. SNP loci were then haplotyped across RAD contigs, using the script rad_haplotyper.pl. The script removed any remaining indel loci, loci with more than five individuals with too few or too many haplotypes given the SNP genotypes, and loci with <95% successfully haplotyped individuals. The resulting GENEPOP file contained one diploid genotype per individual for each remaining RAD contig.

The necessity of sequencing individuals across multiple RAD libraries and HiSeq lanes can result in problematic SNP loci that differ in allele frequencies due to systematic bias in the library preparation and sequencing process (Meirmans *et al.*, 2015). While stringent bioinformatic processing is intended to remove such loci, additional *ad hoc* testing for loci exhibiting effects caused by library organization is often necessary to ensure that the data are free from loci exhibiting such bias. One strategy for identifying problematic loci is to plot the global F_{ST} of each locus as a function of expected heterozygosity and manually inspecting loci that appear as outliers in the distribution. The program LOSITAN (Antao *et al.*, 2008) was used to accomplish this, and the exercise repeated with various combinations of sample localities. As a second assessment, a principal components analysis (PCA), implemented in the package adegenet (Jombart *et al.*, 2008) in R (R Core Team 2015), was used to visualize the data, with individuals coded by RAD library rather than by sample locality.

Assessment of Temporal Stability and Geographic Patterns

Temporal stability of allele frequencies was assessed by comparing the two pairs of samples obtained from the same localities in different years (LLM 2008 vs. LLM 2014 and MAT 2008 vs. MAT 2015). Pairwise F_{ST} was calculated in ARLEQUIN (Excoffier and Lischer, 2010) and significance assessed by a permutation test, with 10,000 permutations. Pairwise F_{ST} did not differ significantly between years at either location: LLM (F_{ST} = -0.00017, P = 0.767) and MAT (F_{ST} = 0.00048, P = 0.179). Inspection of the PCA (Supplemental Figure B1) also confirmed the similarity of the temporal

samples. The 2008 samples from LLM and MAT were then removed for the remainder of the analyses. In addition to establishing temporal stability, the PCA also revealed three distinct clusters of sample localities, corresponding to samples from the western Gulf (LLM, MAT, SAB, and MIS), the eastern Gulf (APA, CEK, and CHA), and the Atlantic (IND, HAR, WAS, and SCA). These regional groupings (western Gulf: WG, eastern Gulf: EG, and Atlantic: ATL) were used in subsequent hierarchical analyses of population structure.

Outlier Detection

The data were screened for the presence of possible loci under selection, using three F_{ST} outlier-detection methods. The first approach used the program LOSITAN, which implements the *fdist* method of outlier detection (Beaumont and Nichols, 1996) and employs coalescent simulations under a neutral island model to identify loci with F_{ST} values that are either higher or lower than expected, given the observed heterozygosity. LOSITAN was run with 100,000 simulations and a false-discovery rate (FDR; Benjamini and Hochberg, 1995) of 0.05. The second approach employed a modified version of the *fdist* method, implemented in ARLEQUIN, which accounts for hierarchical population structure. Localities were grouped into the three regions (ATL, EG, and WG) revealed by PCA, and the analysis run with 50,000 simulations, with an FDR of 0.05. The last approach employed the program BAYESCAN (Foll and Gaggiotti, 2008), which uses a Bayesian approach to estimate the posterior probability that each locus is under selection by comparing models that either incorporate or exclude the effects of selection.

BAYESCAN was run with prior odds of selection relative to drift of 10:1 and a FDR of 0.05. Because loci with low minor allele frequencies can bias results of genome scans (Roesti *et al.*, 2012), loci with a global, major-allele frequency above 0.95 were excluded from all three outlier-detection approaches. The dataset was then split into 'neutral' and 'outlier' components. The final outlier dataset consisted of all loci that had been identified as outliers under directional selection by at least one of the three approaches; the neutral dataset consisted of all remaining loci. All loci detected as outliers due to balancing selection had negative *Fst* values, suggesting that overall mean F_{ST} in the dataset was too low for reliable detection of balancing selection (Beaumont and Balding, 2004; Narum and Hess, 2011). For this reason balancing selection 'outliers' were considered to be neutral loci.

Analysis of Population Structure

Hierarchical analysis of molecular variance (AMOVA) was performed in ARLEQUIN, using regional groupings of localities and significance of variance components assessed by permutation test, with 10,000 permutations. Pairwise F_{ST} indices between individual localities and between regions also were estimated using ARLEQUIN, with significance assessed by permutation, as above. Existence of an isolation-by-distance effect was assessed using Mantel tests as implemented in the *vegan* package (Oksanen *et al.*, 2015) in R, using a matrix of pairwise F_{ST} values, coded as $F_{ST}/(1 - F_{ST})$, and a pairwise matrix of approximate coastline, linear geographic distance (Supplemental Table B1). Pairwise

F_{ST}, AMOVA, Mantel tests, and PCA also were conducted separately with neutral and outlier datasets.

Redundancy Analysis

Redundancy analysis (Meirmans, 2015) was employed to investigate the influence of geographic distance and environmental differences on patterns of observed genetic variation. Redundancy analysis (RDA) combines PCA or principal co-ordinates analysis (PCoA) and multiple regression to measure the influence of a matrix of potential explanatory variables on a matrix of independent variables and has been used in a population-genetics context by Orsini *et al.*, (2012) and Vangestel *et al.*, (2012). Geographic distance between sample localities was coded as a one-dimensional vector of approximate, linear coastline distance (Supplemental Table A1). Environmental data for each locality was obtained from the National Estuarine Eutrophication Assessment database (Bricker *et al.*, 2007). In three cases (WAS, HAR, and CEK), data were not available for the particular bay or estuary sampled; data for the nearest bay or estuary were used in these cases. In all cases, the substituted site (Ossabaw Sound for WAS, Altamaha River for HAR, and Suwannee River for CEK) was less than 20 km from the original sampling locality. A set of 49 environmental variables (Supplemental Table 2) was downloaded and standardized to avoid bias caused by unequal variances. Standardization involved centering each variable value by subtracting the among-locality mean from each value and scaling the centered variables by dividing each by the standard deviation of values among localities. Analysis was performed in R, based on a

modified version (available upon request) of the method presented in Meirmans (2015). Genetic data were transformed into a set of synthetic variables describing the among-locality genetic variation, using PCoA as implemented in the package adegenet. The ten largest principal coordinate axes were retained and used as dependent variables in the RDA model. Linear geographic distances (Supplemental Table A1) were converted to a matrix of polynomials, using the *poly* function in R. To avoid overfitting the model due to correlation of independent variables, forward selection of geographic and environmental variables was conducted with the *ordistep* function in the *vegan* package. The *ordistep* function iteratively adds and drops variables from the model and assesses significance of the change by permutation test in order to select the variables that best explain the data. This was applied to the geographic and environmental variables separately. Following forward selection of variables, the RDA analysis was conducted with the *rda* function (*vegan*). The *varpart* function (*vegan*) was used to partition the total genetic variance into components explained by geography, environment, and the 'shared' component of variance, which represents the variance that is explained by both geography and environment but cannot be decomposed into one or the other due to correlation between the two. Significance of the overall model and of each 'testable' variance component was conducted with the *anova.cca* function (*vegan*), using 1,000 permutations. While variance components attributable to independent variables alone (geography or environment, in this case) are testable under an RDA framework, the shared component of variance is not because it can only be estimated from the other components and thus has zero degrees of freedom (Borcard *et al.*, 2011). The RDA

analysis was conducted separately with datasets containing all loci, only neutral loci, and only outlier loci.

Detection of Genomic Regions under Selection

Global F_{ST} for each locus was calculated, using Weir and Cockerham's (1984) estimator, as implemented in the *pegas* package (Paradis *et al.*, 2010) in R. F_{ST} values for loci that had been successfully incorporated into the linkage map were then plotted against genomic position, using ggplot2 (Wickham *et al.*, 2009). A set of high confidence outliers was identified by selecting loci that met at least one of the following criteria: i) the locus was identified by all three outlier-detection approaches, ii) the locus was within two cM of another outlier, and iii) the locus had a global F_{ST} of at least 0.1.

Identification of Candidate Genes

To develop a list of candidate genes potentially under the influence of selection, a PCA of outlier loci (see Results) was used to separate the loci into two 'functional' groups: PC1 (x-axis) placed ATL in an intermediate position to WG and EG, while PC2 (y-axis) separated regions in the Gulf from the Atlantic. Using the PCA loadings, outlier loci were separated into two groups based on whether they contributed more to PC1 or PC2, respectively. The reference sequences for the both lists of loci were compared to the annotated draft genome of the large yellow croaker (*Larimichthys crocea*), the closest relative to red drum for which a genome sequence was available (Wu *et al.*, 2014). The RefSeq (Pruitt *et al.*, 2007) assembly of the large yellow croaker genome (GenBank

accession: ASM74293v1) was downloaded, along with information regarding the genomic positions of annotated proteins. The croaker genome was converted to a BLAST database, using the *makeblastdb* function in NCBI's Standalone BLAST package (Camacho *et al.*, 2009). The blastn algorithm was applied to match loci to the croaker genome. Loci that mapped to more than one scaffold of the large yellow croaker genome were discarded. For each remaining match, annotated genes that completely or partially overlapped the region defined by 100 kb before the start position and after the end position of the match were selected as candidate genes.

RESULTS

Reduced Representation Reference Genome Construction

After assembly, the reduced-representation, reference genome contained 38,887 RAD contigs, with a mean size of 264.9 bp. Additional filtering for contigs below minimum threshold size (<150 bp) and containing rRNA sequences resulted in exclusion of 4,848 and 174 contigs, yielding a final reference assembly containing 33,865 RAD contigs, with mean size of 284.6 bp. Assuming a total red drum genome size of ~ 810 MB (Gold *et al.*, 1988), the total reference sequence length of 9,638,003 bp covered approximately 1.19 percent of the entire genome.

Linkage Map

Sample Collection and Genotyping

After filtering, the dataset for Family A consisted of 786 RAD contigs (containing 1,383 SNPs) and 72 individual progeny. The dataset for Family B consisted of 1,340 RAD contigs (containing 2,620 SNPs) and 81 individual progeny. The final number of usable RAD contigs differed considerably between families (786 vs. 1,340), largely as a result of differences in overall tissue (DNA) quality of samples between the two families. After combining the RAD contig data with microsatellite and EST-SSR genotypes assayed previously from the same individuals (Portnoy *et al.*, 2010, Hollenbeck *et al.*, 2015), the total mapping dataset consisted of 1,218 and 1,779 loci for Families A and B, respectively.

Linkage Mapping

The consensus linkage map (Supplemental Figure B2) contained a total of 2,275 loci, including 1,794 RAD contigs (consisting of 3,462 SNP loci), 437 anonymous microsatellite loci, and 44 EST-SSRs. The mean number of loci per linkage group was 94.79 and the mean marker interval was 0.94 cM. The average length of a linkage group and the total map length were 87.72 cM and 2,105.30 cM, respectively. However, because of the tendency of MERGEMAP to inflate the total size of linkage groups when combining maps (Kahn *et al.*, 2012), the average of individual-specific maps may represent a more accurate estimate of total map size. In this case, the mean length of

Table 4.1 Summary statistics for red drum linkage maps. Column names represent maps constructed using various subsets of mapping individuals: Consensus is consensus map with all individuals; Family A is family-specific map for Family A (male and female); Family B is family-specific map for Family B (male and female); AF is Family A female; AM is Family A male; BF is Family B female; and BM is Family B male.

	Consensus	Family A	Family B	AF	AM	BF	BM
Total Loci	2275	1001	1569	678	674	961	976
Msat Loci	437	327	334	228	241	194	190
EST-SSR Loci	44	33	32	26	25	23	20
RAD Loci	1794	641	1203	424	408	744	766
SNP Loci	3462	1170	2456	847	884	1693	1768
Mean Linkage Group Size (cM)	87.72	77.74	70.08	76.08	74.10	73.42	63.63
Mean Loci per Linkage Group	94.79	41.71	65.38	29.48	28.08	40.04	40.67
Mean Marker Interval (cM)	0.94	1.91	1.09	2.64	2.73	1.87	1.59
Total Map Length (cM)	2105.30	1865.65	1682.01	1749.84	1778.44	1762.12	1527.22

linkage groups and the total map length, averaged across individual-specific maps, was 71.81 cM and 1,704.40 cM, respectively. Summary statistics for the consensus map and family- and individual-specific maps are presented in Table 4.1.

Population Genomics

A summary of data filtering procedures, including number of sites, contigs, and individuals excluded at each filtering step, can be found in Supplemental Table A3. The final dataset consisted of genotypes for 462 individuals and 1,539 (haplotyped) RAD contigs, consisting of 2,860 SNP loci.

Outlier Detection

Prior to outlier detection, 20 loci with a major allele frequency > 0.95 were excluded from the dataset for all analyses. Following application of an FDR of 0.05, the three outlier-detection methods identified a total of 146 outliers (9.49% of all loci) putatively under directional selection and 37 outliers (2.40% of all loci) putatively under balancing selection. Of the three outlier-detection methods, LOSITAN was the least conservative (142 directional, 37 balancing), BAYESCAN was intermediate (100 directional, 4 balancing), and ARLEQUIN was the most conservative (52 directional, 10 balancing). Of outliers under directional selection, 96 loci were classified as outliers by at least two methods; 52 loci were classified as outliers by all three methods. All 52 loci classified as outliers by ARLEQUIN also were classified as outliers by the other two methods. The distribution of F_{ST} values and expected heterozygosity for each locus, along with the

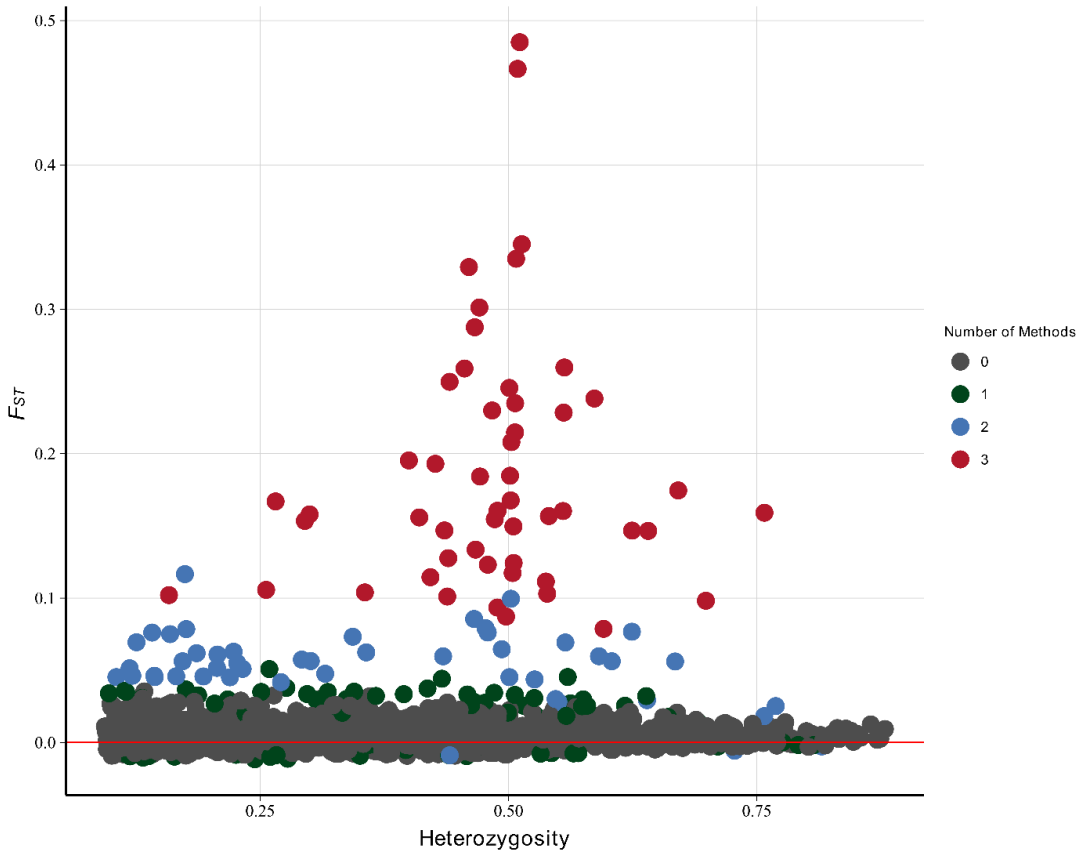


Figure 4.2 F_{ST} and expected heterozygosity for each locus. The plot shows F_{ST} (y-axis) and expected heterozygosity (x-axis) for each locus in the dataset. Each point represents a locus, and the color of each point indicates the number of outlier detection methods that determined the locus to be an outlier.

number of methods for which each locus was determined to be an outlier, is presented in Figure 4.2). The final neutral and outlier datasets consisted of 1,393 and 146 loci, respectively.

Analysis of Population Structure

Inspection of the PCA (Figure 4.3a) for all loci revealed three distinct clusters representing three geographic regions: WG, EG, and ATL. The two largest principal components explained 1.04 percent and 0.71 percent of the variance, respectively. Hierarchical AMOVA (Table 4.2) revealed significant differences among regions and among localities within regions, with among-region and among-localities (within region) differences accounting for 1.37 percent and 0.09 percent of the total genetic variation, respectively. Pairwise F_{ST} analysis (Figure 4.3b; Supplemental Table A4) reinforced the pattern of significant genetic differences among regions, as all pairwise comparisons between localities in different regions differed significantly following FDR correction; significant pairwise comparisons between localities within regions included APA and CHA (both in the EG), and IND which differed significantly from all other localities in ATL. F_{ST} values between WG and EG generally were larger than F_{ST} values between WG and ATL (Figure 4.3b). Pairwise estimates of F_{ST} between regions revealed that the F_{ST} for WG vs. EG (0.021, $P = 0.000$) was greater than that of WG vs. ATL (0.012, $P = 0.000$) and EG vs. ATL (0.010, $P = 0.000$). The Mantel test revealed a significant relationship between geographic distance and F_{ST} ($r = 0.505$, *P-value* = 0.008; Figure 4.3c).

Table 4.2 Hierarchical analysis of molecular variance. Hierarchical analysis of molecular variance (AMOVA) was based on datasets containing all loci (n = 1,539), only neutral loci (n = 1,393), and only outlier loci (n = 146). df is degrees of freedom; F is fixation index; P is probability that F = 0. Significant values of F are in bold. Significance was assessed by permutation test with 10,000 permutations.

Level	df	Sum of Squares	Variance Component	% Variation	*F*	*P*
All Loci						
Among regions	2	2854.133	4.17574	1.37	**0.01372**	0
Among localities within regions	8	2553.381	0.27101	0.09	**0.0009**	0.00178
Within localities	789	236666.979	299.95815	98.54	**0.01461**	0
Neutral Loci						
Among regions	2	1158.502	1.14044	0.42	**0.00417**	0.0003
Among localities within regions	8	2214.052	0.06142	0.02	0.00023	0.8902
Within localities	789	214924.738	272.40144	99.56	**0.00439**	0
Outlier Loci						
Among regions	2	1695.632	3.03529	9.85	**0.09854**	0.0001
Among localities within regions	8	339.329	0.20958	0.68	**0.00755**	0
Within localities	789	21742.241	27.55671	89.47	**0.10535**	0

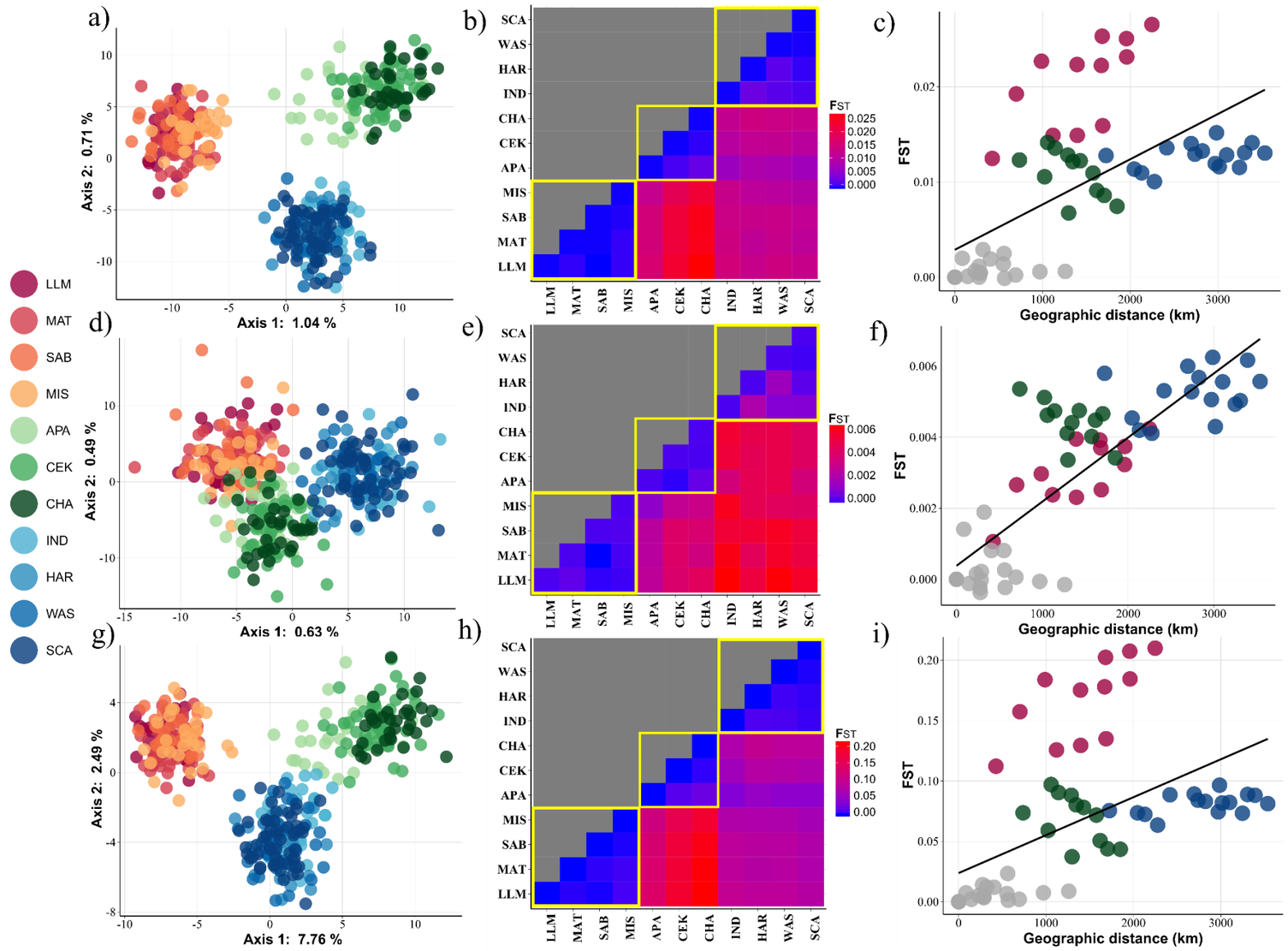


Figure 4.3 Population structure of red drum. Population structure was evaluated with datasets consisting of all loci (n = 1,539; panels a-c), only neutral loci (n = 1,393; panels d-f), and only outlier loci (n = 146; panels g-i). Column 1 (panels a, d, and g) shows a principal components analysis (PCA), using each dataset. Column 2 (panels b, e, and h) shows a heatmap of pairwise F_{ST} estimates for each locality, using each dataset; blue squares represent low F_{ST} values and red squares represent high F_{ST} values between localities. Localities (left to right) are organized geographically (west to east) along the coastline, and putative regional groups (western Gulf, eastern Gulf, and Atlantic) are highlighted with yellow boxes. Column 3 (panels c, f, and i) shows the relationship between genetic distance (F_{ST}; y-axis) and geographic distance (km; x-axis) for each pair of localities, using each dataset. Pairings between localities in different regions are highlighted by different colors (within region comparisons = grey; eastern Gulf vs. Atlantic = green; western Gulf vs. eastern Gulf = red; western Gulf vs. Atlantic = blue). The black line in each panel shows the best linear fit for the relationship.

Inspection of the PCA for neutral loci (Figure 4.3d) also revealed three clusters corresponding to WG, EG, and ATL. The two largest principal components explained 0.63 percent and 0.49 percent of the total genetic variance, respectively. EG was placed between WG and ATL along PC1 (x-axis), whereas WG and ATL were placed closer together along PC2 (y-axis). Hierarchical AMOVA (Table 4.2), based on the neutral dataset, revealed significant among-region divergence (F_{CT} = 0.0041, P = 0.000), but not among-localities within regions (F_{SC} = 0.0002, P = 0.890). Among-region genetic variation accounted for 0.42 percent of the total genetic variation. F_{ST} estimates (Supplemental Table A4) for all pairwise comparisons between localities in different regions differed significantly from zero except for MIS vs. APA; the relative magnitude of F_{ST} values between localities was consistent with the geographic distribution of regions, with localities in WG and ATL having the highest pairwise F_{ST} values (Figure 4.3e). F_{ST} estimates for all pairwise comparisons between localities within regions except for IND vs. HAR were non-significant. Estimates of pairwise F_{ST} between regional groupings confirmed the overall pattern; WG vs. ATL was the most divergent (F_{ST} = 0.005, P = 0.000), followed by EG vs. ATL (F_{ST} = 0.004, P = 0.000), and WG vs. EG (F_{ST} = 0.003, P = 0.000). The Mantel test based on neutral loci (Figure 4.3f) revealed a strong, significant relationship between geographic and genetic distance (r = 0.791, P = 0.001).

PCA for outlier loci (Figure 4.3g) similarly revealed three regional clusters. The two largest principal components explained 7.76 and 2.49 percent of the total genetic variation, respectively. ATL was placed in an intermediate position between WG and

EG along PC1 (x-axis), whereas WG and EG were placed closer together along PC2 (y-axis). Further, the orientation of EG localities relative to ATL was discordant with respect to the geographic relationship of the samples; APA, which is the EG locality most distant from ATL by coastal distance, was most similar to ATL (Figure 4.3g). To test for the presence of a latitudinal effect on allele frequencies, association of allele frequencies to latitude in EG and ATL was assessed via linear regression. Thirty-three outlier loci had alleles that were significantly associated with latitude among localities in the two regions ($P < 0.05$). Based on PCA using only EG and ATL localities (Supplemental Figure B3), the 33 loci explained 21.79% of the variance along the primary PC axis (PC1). Hierarchical AMOVA (Table 4.2) revealed significant differences among regions ($F_{CT} = 0.0985$, $P = 0.000$) and among localities within regions ($F_{SC} = 0.0075$, $P = 0.000$). The proportion of genetic variation explained by among-region differences and among-localities within region differences was 9.85 and 0.68 percent, respectively. F_{ST} estimates (Supplemental Table A4) for all pairwise comparisons between localities in different regions differed significantly from zero; F_{ST} values between localities in WG and those in EG generally were larger than F_{ST} values between localities in WG and ATL (Figure 4.3h). F_{ST} estimates for pairwise comparisons between localities within regions also differed significantly from zero except for LLM vs. SAB, SAB vs. MIS, and WAS vs. SCA. Pairwise F_{ST} of regional groupings indicated a similar pattern; the estimate of F_{ST} between WG and EG ($F_{ST} = 0.168$, $P = 0.000$) was over twice that of WG vs. ATL ($F_{ST} = 0.076$, $P = 0.000$) and EG vs. ATL ($F_{ST} = 0.063$, $P = 0.000$).

Table 4.3 Redundancy analysis. Redundancy analysis (RDA) was based on all loci (n = 1,539), neutral loci (n = 1,393), and outlier loci (n = 146). $R^2_{Adj.}$ is proportion of total among-locality genetic variation explained by all of the independent variables (geography + environment + shared); Geography is proportion of total among-locality genetic variation explained by geography alone; Environment is proportion of total among-locality genetic variation explained by environmental variables alone; Shared is proportion of total among-locality genetic variation explained by the combination of geography and environmental variables; Residual is proportion of total among-locality genetic variation not explained by the model. Significant values ($P < 0.05$) are in bold.

	All Loci	Neutral	Outlier
Total Adjusted R^2	**0.252**	**0.126**	**0.691**
Geography	0.022	0.016	0.043
Environment	0.062	0.017	**0.219**
Shared	0.168	0.093	0.429
Residual	0.748	0.874	0.309

The Mantel test (Figure 4.3i) revealed a weak, but significant relationship between genetic and geographic distance ($r = 0.357$, $P = 0.033$).

Redundancy Analysis

Forward selection of geographic distance variables in polynomial form resulted in selection of first- and second-order polynomials in the model; forward selection of environmental variables resulted in selection of three environmental variables: mean oceanic (outside of the estuary) concentration of dissolved inorganic phosphates (oceanic DIP), minimum oceanic salinity, and average wind speed. A summary of results of redundancy analysis for all loci, only neutral loci, and only outlier loci is given in Table 4.3. For all loci, the set of independent (constraining) variables, which included both geographic and environmental variables, explained a significant proportion of the total among-locality variance ($R^2_{Adj} = 0.252$, $P = 0.003$). Neither geography nor environment alone explained a significant component of among-locality variance, although the component of shared variance, while not testable for significance (see Methods), was relatively large. For neutral loci, the set of independent variables explained a significant component of among-locality genetic variance ($R^2_{Adj} = 0.126$, $P = 0.007$), but neither geography nor environment alone were significant. For outlier loci, the set of independent variables explained a large and significant proportion of among-locality genetic variance ($R^2_{Adj} = 0.691$, $P = 0.009$). However, the component of variance attributable to geography was non-significant ($R^2_{Adj.} – 0.043$, $P = 0.079$), whereas the component attributable to environment was significant ($R^2_{Adj} = 0.219$, $P =$

0.015). Inspection of the RDA eigenvalues computed with only outlier loci revealed that the constrained genetic variance (the variance attributable to the independent variables) was largely explained by the two largest RDA axes, which accounted for 69.03 and 20.06 percent of the constrained genetic variance. An RDA biplot for outlier loci (Figure 4.4) reveals the genetic relationships among localities overlain with vectors of environmental variables. Oceanic DIP appears to be highest in the WG and lowest in the EG, minimum oceanic salinity appears lowest in the WG and highest in the ATL, and average wind speed appears highest in the ATL and lowest in the EG.

Detection of Genomic Regions under Selection

Of 1,539 RAD loci in the complete dataset, 746 (48.5%) were segregating in at least one mapping cross and were successfully added to the linkage map. Of 746 mapped loci, 73 were identified as outliers by at least one outlier-detection method. Based on criteria noted above, i.e., identified as an outlier by all three methods, within two cM of another outlier, or with a global $F_{ST} > 0.1$, 45 loci were designated as high-confidence outliers. These outliers tended to be grouped non-randomly on linkage groups (Supplemental Figure B4); 39 of the 45 high-confidence outlier loci were located within two cM of another outlier locus and grouped into 15 clusters of outlier loci on ten different linkage groups (Figure 4.5).

The first two principal components of the PCA of outlier loci (Figure 4.3g) explained a majority of detected, among-region genetic variation; PC1 separated all three regions, with ATL localities positioned between WG and EG localities, while PC2

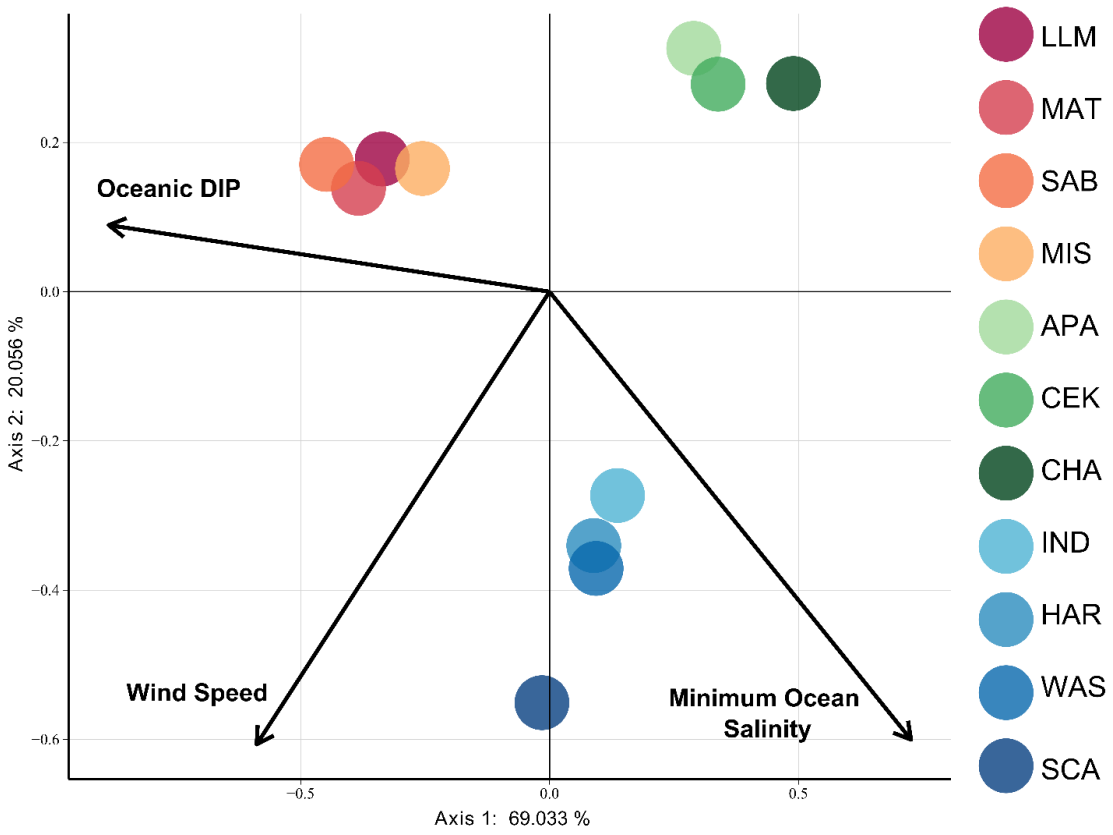


Figure 4.4 A biplot from redundancy analysis. Colored points are represent localities and positions are indicative of the genetic relationships between localities, generated with principal coordinates analysis (PCoA), using the outlier dataset. Arrows represent the three environmental variables used in the redundancy analysis (RDA) model, and their directionality indicates the vector upon which each variable correlates with the genetic relationship among localities. Oceanic DIP = annual mean oceanic (outside of the estuary) concentration of dissolved inorganic phosphates; Wind Speed = average estuary wind speed; Minimum Ocean Salinity = annual minimum oceanic (outside of the estuary) salinity.

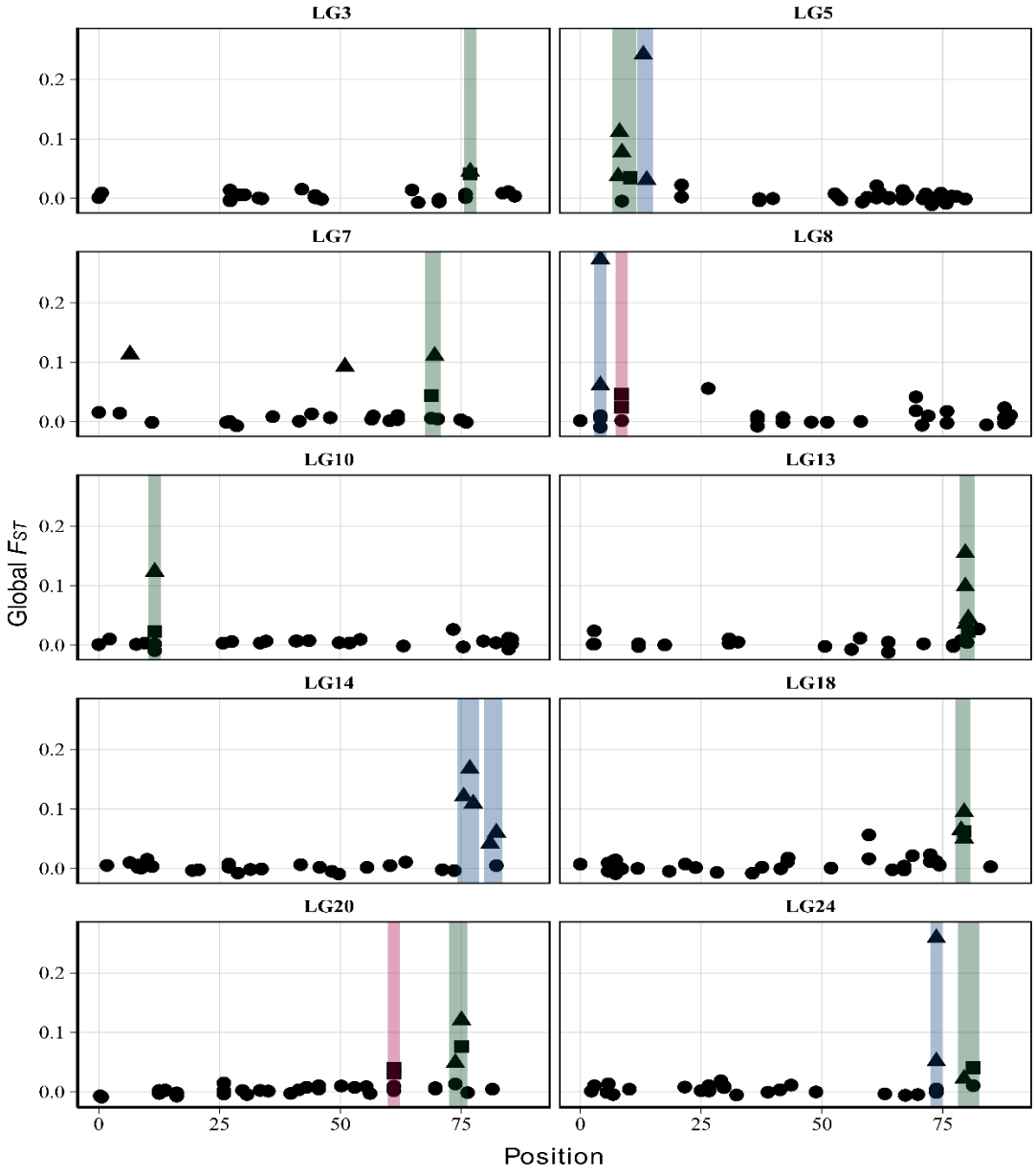


Figure 4.5 Location of 15 outlier clusters in the red drum genome. Outlier clusters are regions containing at least two high-confidence outliers spaced less than 2 centiMorgans (cM) apart. Each panel represents one of the ten linkage groups where outlier clusters were identified and shows the global F_{ST} for each locus (y-axis) against position on the linkage group (cM; x-axis). The shape of each point represents the type of locus: triangles = loci which contributed a larger proportion to the x-axis of the outlier PCA (x-loci); squares = loci which contributed a larger proportion to the y-axis of the outlier PCA (y-loci); circles = not outliers. Outlier clusters are highlighted with colored rectangles: blue rectangles = outlier clusters containing only x-loci; red rectangles = outlier clusters containing only y-loci; green rectangles = outlier clusters containing both x- and y-loci.

separated ATL localities from both WG and EG localities. Of the 45 high-confidence outlier loci, 32 contributed to a larger proportion of the variance along PC1, while the remaining 13 loci contributed a larger proportion of the variance along PC2. Of the 15 clusters of outlier loci identified on the linkage map, five contained loci contributing to PC1, two clusters contained loci contributing to PC2, and eight clusters contained loci contributing to both PC1 and PC2.

Identification of Candidate Genes

The BLAST search of 32 outlier loci contributing to PC1 resulted in 20 strong matches to the large yellow croaker genome assembly. The regions immediately surrounding each of the 20 hits (± 100 kb) contained 59 candidate genes. The BLAST search of 13 outlier loci contributing to PC2 resulted in nine matches. The regions surrounding each of the nine hits contained 35 candidate genes. A summary of these results is presented in Supplemental Table A5.

DISCUSSION

Linkage Map

In total, 1,794 haplotyped RAD contigs, consisting of 3,462 SNP loci, were added to the red drum linkage map. Addition of these loci reduced the mean marker interval to less than one cM. The total length of the consensus map (2,105.30 cM) is larger than reported previously for microsatellite-based maps (1196.9 cM, Portnoy *et al.,* 2010; 1815.3, Hollenbeck *et al.,* 2015). Much of this discrepancy is probably a result of

combining maps from different families, using MERGEMAP, which has been reported to inflate total map length (Khan *et al.*, 2012). Relative to the average total length of individual-based maps reported in this study, the total length of the consensus map was inflated by ~24 percent, consistent with the results of McKinney *et al.,* (2015) who found that MERGEMAP increased the size of the Chinook salmon consensus map by ~30 percent per additional family. However, there exists a tradeoff between accuracy of map lengths and inclusion of loci on the map (Chapter 2 of this dissertation). Because the purpose of the map in this study was to determine the relative positions of loci rather than the frequency of recombination between them, addition of more loci to the consensus map by combining information between mapping families at the expense of more accurate map distances is acceptable. Overall, the dense linkage map will be a valuable resource for future studies of red drum, particularly for genetic improvement in commercial aquaculture where linkage maps are critical tools for QTL mapping and marker-assisted selection (Liu and Cordes, 2004).

Population Genomics

Temporal Stability

Temporal stability of allele frequencies was documented by comparing samples collected at two localities, LLM and MAT, in spring 2008 and winter 2014/2015. The >6 year sampling interval is sufficiently long to ensure that the temporal samples contained different juvenile cohorts, given that juvenile red drum typically inhabit natal bays and estuaries for 3-5 years before becoming reproductively mature and joining the

offshore adult population (Pattillo *et al.,* 1997). While temporal samples were obtained from these two localities only, the results provide additional evidence that observed genetic differences among geographic localities reflect relatively stable, long-term patterns and are not caused by year-to-year variability in juvenile recruitment (Scharf, 2000, Gold and Turner, 2002).

Outlier Detection

The frequency of outliers (~9.5% of all loci analyzed) putatively under directional selection is consistent with results of a meta-analysis of 20 genome-scan studies by Nosil *et al.,* (2009), wherein the average per cent of outlier loci was ~8.5. Nevertheless, the total number of outliers observed in this study almost certainly represents an overestimate due to false positives arising from a number of factors (see Meirmans *et al.,* 2015 for discussion). One potential source of false positives in genome scans is the presence of hierarchical structure among sample localities (Excoffier *et al.,* 2009). This is because when localities are structured hierarchically, individual localities do not represent independent pools of migrants, as is assumed under a typical island model (Wright, 1943), and the overall variance in F_{ST} will be inflated if structure is not accounted for in the null distribution of F_{ST} (Hermisson, 2009, Excoffier *et al.,*2009). This is a particularly important consideration for red drum given the hierarchical structure observed in this study. Consistent with this, the outlier detection algorithm employed in ARLEQUIN, which accounts for the effects of predefined hierarchical structure and was the most conservative method applied, identified 52 outliers under

directional selection (52 loci; 3.4% of all loci assayed). Notably, the other two approaches also identified the same 52 outliers. We used all 146 identified outliers when defining neutral and outlier datasets in this study because a few outlier loci can dramatically influence patterns of variation if not removed from a 'neutral' dataset (Luikart *et al.*, 2003).

Population Genetic Structure

Population-genetic analyses of neutral, outlier, and full (all loci) datasets revealed the existence of three, genetically distinct populations of red drum, where sample localities were grouped into WG, EG, and ATL regions. The similarity in patterns of genetic variation and divergence between the full dataset and that containing F_{ST} outliers demonstrates that outlier loci strongly influenced the recovered pattern of population structure observed when all loci were analyzed together. Because the patterns exhibited by the full dataset were so heavily influenced by the outlier loci, the remainder of the discussion will focus on only the neutral and outlier datasets.

Genetic divergence among neutral loci was significant in pairwise comparisons between the three regions (ATL, EG, and WG), indicating existence of contemporary and/or historical barriers to gene flow. The significant genetic differences between red drum in the Gulf and Atlantic has been reported previously (Gold and Richardson, 1991; Gold *et al.*,1993, 1999; Seyoum *et al.*,2000; Chapman *et al.*,2002), as have genetic differences between populations in the Gulf and Atlantic in a variety of other coastal marine species (Supplemental Table A6). The marine fishes that differ genetically on

either side of the Florida peninsula include a broad range of taxonomic groups and general life histories (Supplemental Table A6), suggesting that both historical and contemporaneous barriers to gene flow likely occur around peninsular Florida (Reeb and Avise, 1990; Gold *et al.,* 1993; Gold and Richardson, 1998; Seyoum *et al.,* 2000; Portnoy *et al.,* 2015).

The significant genetic discontinuity between red drum in the WG and EG adds to the list of other coastal marine species where genetic differences between the WG and EG regions has been reported (Supplemental Table A6). There also are morphological and/or genetic data for more than 15 pairs of sister taxa that display a similar WG/EG distribution (reviewed in Portnoy and Gold, 2012). Distributional data for the pairs of sister taxa (Figure 1 and Appendix S1 in Portnoy and Gold, 2012) indicated that the boundary between the regions was broad and in the area between longitude 84 W and 89 W. The boundary has been referred to as a vicariant (Dahlberg, 1970; McClure and McEachran, 1992) or marine suture-zone (Portnoy and Gold, 2012). Because of the proximity of the zone to the Mississippi River, hypotheses proposed to account for the zone include physical and/or ecological barriers to gene flow stemming from the river's outflow at different points in the past 2-3 million years (Karlsson *et al.*, 2009; Portnoy and Gold, 2012; Portnoy *et al.*, 2014). The same or other barriers also could restrict contemporary gene flow. Additional contemporary barriers might include the narrowing of the continental shelf around DeSoto Canyon at ~85 W (Johnson *et al.,* 2009) and/or the Loop Current (Karlsson *et al.,* 2009) which periodically forms an intense clockwise

flow that can reach as high as the Mississippi river delta and the continental Florida shelf (Huh *et al.*, 1981; Wiseman and Dinnel, 1988).

Tests of genetic homogeneity, using neutral loci, among sample localities within the three regions were non-significant except for the comparison of IND versus HAR in the Atlantic. However, there also was a strong and highly significant isolation-by-distance effect ($r^2 = 0.626$), based on neutral loci, consistent with prior genetic studies of red drum in the Gulf (Gold and Richardson 1991; Gold *et al.*,1993, 1999) and the model proposed by Gold *et al.,* (2001), in which gene flow occurs primarily (but not exclusively) between adjacent bays and estuaries distributed linearly along the coastline and where the probability of gene exchange decreases with increasing geographic distance. Interestingly, the test of pairwise genetic homogeneity between the samples from MIS in the WG and APA in the EG was non-significant, whereas all other tests of homogeneity between samples from different regions were significant. The two localities (MIS and APA) are only 427 km from one another, well within the distances between most localities within regions. This observation suggests that periodic gene flow still occurs between the two localities.

Genome scan approaches are a powerful tool for detecting loci under selection in natural populations; however, interpretation of the results of such strategies must be done with care, as non-selective forces such as demographic change can mimic the effects of selection in the genome (Teshima *et al.,* 2006; Lotterhos *et al.,* 2014; Poh *et al.,* 2014). Because of this, it is important to consider other sources of evidence for selection (Storz, 2005). In this study, there are three lines of evidence indicating that

natural selection has played an important role in shaping the observed patterns of genetic variation in red drum. These are: 1) differences in the patterns of divergence between regions based on neutral and putatively adaptive variation, 2) correlations between adaptive genetic variation and environmental factors, and 3) the presence of outlier loci which are non-randomly distributed in the genome.

Genetic divergence among outlier loci also was significant among regions but differed from genetic patterns indicated by neutral loci in that significant divergence also was found among localities within regions. In addition and also in contrast to genetic patterns indicated by neutral loci, the degree of divergence between WG and EG among outlier loci was more than two-fold greater than that between EG and ATL and WG and ATL. The same geographically discordant pattern also was indicated by the location of the three regional clusters along the primary axis (PC1) of the PCA analysis. A second pattern indicated by the outlier loci, displayed on the second PCA axis (PC2) was a separation of Gulf localities from Atlantic localities, which suggests that divergent selection has operated between the two regions. Further, a considerable proportion of the genetic variation explained by outlier loci between EG and ATL was correlated with latitude, a pattern that has been documented previously among bonnethead sharks in the same region (Portnoy *et al.*, 2015). Latitude-related genetic differences in populations of marine fishes have been attributed to temperature-dependent growth rates (Conover and Present, 1990) and differential parasite loads (Poulin and Morand, 2000), although other adaptive differences related to the transition from temperate to sub-tropical climate in southern Florida cannot be ruled out (Portnoy *et al.*, 2015). Overall, because the outlier

loci grouped the localities by region despite geographic discordance, it would appear that local adaptations may be driven largely by environmental factors that vary on a regional scale rather than at the level of individual bays and estuaries.

There also was evidence of significant heterogeneity, based on outlier loci, among localities within regions; significant F_{ST} values were found in all but three pairwise comparisons between localities in the same region: LLM/SAB and SAB/MIS in the WG and WAS/SCA in the ATL. However, the magnitude of significant F_{ST} values for within-region comparisons were roughly an order of magnitude less than between-region comparisons, suggesting that within-region differences, if any, are likely small. This bears further investigation because existence of temporally stable genetic differences within regions would have important implications for resource allocation and decision-making with regards to management of the red drum fishery.

Redundancy analysis supported the inferences of population structure indicated by neutral and outlier loci. For all three datasets, the combination of geographic and environmental variables explained a significant proportion of the total among-locality variance. Geography alone did not explain a significant proportion of the among-locality genetic variance in any of the datasets; however, environmental variables alone did explain a significant proportion of the among-locality genetic variance for outlier loci ($R^2_{Adj} = 0.219$, $P = 0.015$), further supporting the hypothesis that the patterns of population structure attributed to outlier loci are being driven by natural selection.

The three explanatory environmental variables chosen in model selection were oceanic DIP, minimum oceanic salinity, and average wind speed. All three variables

differ on broad (regional) geographic scales rather than between geographically adjacent bays and estuaries, and all three would be expected to impact habitats where red drum are found. DIP, for example, reflects phosphorus availability and has been shown to correlate with primary autotrophic production in marine and estuarine habitats (Howarth, 1988; Paytan and McLaughlin, 2007) and thus to form the base of the food web and influence species distributions and ecosystem structure (Ryther, 1969). The primary source of oceanic phosphorus is continental weathering and erosion (Paytan and McLaughlin, 2007), suggesting that concentration of DIP is likely correlated with freshwater inflow. Salinity differences can also impact red drum survival, as hatching of red drum under experimental conditions and a range of salinities characteristic of natural conditions have shown that salinity differences significantly influence hatching success and 24-hour survival of larvae (Holt *et al.,* 1981). Finally, wind is a major driver of ocean currents and upwelling (Emerson and Hedges, 2008) and influences survival of fish eggs and larvae (Norcross and Shaw, 1984). This is undoubtedly important for red drum given that adults spawn in nearshore habitats and the transport of larvae into bay and estuarine nurseries is dependent on favorable current conditions (Matlock, 1987; Rooker and Holt, 1997).

The WG is characterized by the highest levels of DIP and lowest minimum oceanic salinity, variables that are likely a function of the large amount of freshwater inflow from the Mississippi and Atchafalaya rivers (Rabalais *et al.,* 1996; Huang *et al.,* 2004) and the scale of anthropogenic phosphorus sources such as agricultural runoff (Paytan and McLaughlin, 2007). Wind speed is typically higher in both the WG and

ATL due to the direction of prevailing winds, which for most of the year move southeast to northwest across the southern U.S. (NREL, 2015). For all three environmental variables, localities within regions were more similar in magnitude, and for oceanic DIP and wind speed, sample localities in WG were more similar to sample localities in ATL than to those in EG; these observations are consistent with the genetic patterns observed for the outlier dataset. It should be noted that because these data are only appropriate for identifying correlations between environmental factors and genetic signal, it is only possible to speculate about potential factors driving natural selection. Clearly, further study is needed to address specific hypotheses about environmental factors important for local adaptation in red drum.

The non-random grouping of outlier loci into clusters on different linkage groups is consistent with theoretical results (Beaumont, 2005) and observations from natural populations (Turner *et al.*, 2005; Bradbury *et al.*, 2010; Hemmer-Hansen *et al.*, 2013) of ongoing or past divergent selection. The patterns of variation in the genome also are consistent with the existence of 'genomic islands of divergence' (Wu, 2001; Turner *et al.*, 2005) which contain loci under divergent selection. It is difficult to ascertain whether individual clusters consist of a single locus under strong selection or multiple loci under selection because the number of outlier RAD loci in a given cluster is likely related to restriction-site frequency, which determines marker density of a given region, and to the local recombination rate, which modulates the extent to which loci in proximity of a specific locus or loci under selection can 'hitchhike' to elevated divergence (Barton, 2000). However, it is predicted that loci of adaptive importance

should tend to aggregate together in the genome because mutations contributing to adaptive divergence tend to accumulate in regions where recombination is reduced (Navarro and Barton, 2003) or that already are under the influence of divergent selection (Nosil *et al.,* 2009). There is evidence that the observed outlier clusters in the red drum genome do contain multiple loci under selection because most outlier clusters contained loci that contributed heavily to the two functional groupings (PC1 and PC2 of the PCA analysis) that contributed to the distinct patterns of geographic divergence. Further, loci under selection also are simultaneously under the influence of other population genetic forces, which makes it difficult to unambiguously assign loci to probable functional groupings. Regardless, the candidate genes identified will be useful for future studies investigating the genetic basis of adaptation in red drum.

Relative Role of Neutral and Adaptive Processes

Separating genetic variation into neutral and adaptive components allowed unique insight into the way that different micro-evolutionary forces have influenced wild populations of red drum. Based on significant heterogeneity in neutral loci between red drum on either side of the Florida Peninsula and a component (PC2 in PCA analysis) of variation in outlier loci that separated red drum in the Gulf from those in the Atlantic, it seems likely that both genetic drift and divergent selection have contributed to the genetically different populations in the Gulf and Atlantic.

The observed genetic differences between the WG and EG present a more interesting problem as heterogeneity in both neutral and outlier loci differences exist yet

there still appears to be gene flow between the eastern-most locality (MIS) sampled in the WG and the western-most locality (APA) in the EG. This suggests the possibility that a contemporary, physical barrier to gene flow does exist between WG and EG and that divergent selection is acting in conjunction with genetic drift to drive genetic divergence between the two regions (Nosil *et al.,* 2009). Possible contemporary barriers to gene flow at the larval stage were noted above (DeSoto Canyon and the invasion of the Loop Current into the EG), but neither likely affect adult and sub-adult red drum movement over the long term given that red drum are long-lived (Ross *et al.,* 1995) and are known to be capable of long-distance (>600 km) movement (Bacheler *et al.,* 2009).

It also is possible is that ecological differences between the WG and EG promote important local adaptations that decrease effective levels of migration due to reduced survival and reproductive success of immigrant individuals; this also would impact divergence in neutral loci. This process is termed isolation-by-adaptation or IBA (Nosil *et al.,* 2008). In addition to the environmental variables described above (oceanic DIP and minimum salinity, wind speed), the WG and EG are characterized by different sediment types: from terrigenous mud and silt in the WG to carbonate sediment in the EG (Wilhelm and Ewing, 1972; Bert, 1986); the two regions also have been identified as separate geological provinces (Uchupi, 1975). On the other hand, the same pattern of divergence (WG vs. EG) has been documented in several, taxonomically diverse species with different life histories, suggesting that a common vicariant event (or events) that influenced them simultaneously, occurred at some point in the past. If so, historical

vicariance could have initially promoted divergence in neutral loci between red drum in the two regions that was then reinforced by local adaptation.

CONCLUSION

Genomics techniques offer unprecedented insight into the micro-evolutionary forces that influence wild populations. In this study, a genomic approach was used to study population structure of red drum in U.S. waters, to examine the relative effects of neutral and adaptive forces in shaping contemporary patterns of genetic variation and population structure, and to explore the genomic distribution of outlier loci. The results of the study have implications for commercial aquaculture of red drum by providing a dense linkage map for future QTL mapping and marker-assisted selection (Liu and Cordes, 2004). By describing the geographic scales of population structure and local adaptation, the study also provides useful information to inform selection of brood fish to mitigate possible genetic risks of escapement (Hindar *et al.,* 1991) from red drum aquaculture facilities. Understanding the geographic scale of population structure and local adaptation also can benefit red drum restoration-enhancement programs (McEachron *et al.,* 1993) by enabling the efficient matching of brood stock to release sites. Finally, characterizing the geographic structuring of populations will enable managers to more efficiently manage the red drum fishery and conserve important adaptive genetic variation, which can be expected to lead to a more sustainable natural resource.

REFERENCES

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805–1814.

Alam M, Han KI, Lee DH, Ha JH, Kim JJ (2012). Estimation of effective population size in the Sapsaree: a Korean native dog (*Canis familiaris*). *Asian-Australasian J Anim Sci* **25**: 1063–72.

Allendorf FW, England PR, Luikart G, Ritchie PA., Ryman N (2008). Genetic effects of harvest on wild animal populations. *Trends Ecol Evol* **23**: 327–337.

Allendorf FW, Hohenlohe PA, Luikart G (2010). Genomics and the future of conservation genetics. *Nat Rev Genet* **11**: 697–709.

Anderson JD, Karel WJ, Mione ACS (2012). Population structure and evolutionary history of southern flounder in the Gulf of Mexico and western Atlantic Ocean. *Trans Am Fish Soc* **141**: 46–55.

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

Antao T, Perez-Figueroa A, Luikart G (2011). Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evol Appl* **4**: 144–154.

ASMFC {Atlantic States Marine Fisheries Commission} (2002). *Interstate Fishery Management Plan for Red Drum, Amendment 2, Number 38*. ASMFC. Washington, D.C.

Bacheler NM, Paramore LM, Burdick SM, Buckel JA, Hightower JE (2009). Variation in movement patterns of red drum (*Sciaenops ocellatus*) inferred from conventional tagging and ultrasonic telemetry. *Fish Bull* **107**: 405–419.

Baird NA Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, *et al.* (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**: e3376.

Balloux F (2004). Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution* **58**: 1891–1900.

Barton NH (2000). Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci* **355**: 1553–1562.

Beaumont MA (2005). Adaptation and speciation: what can FST tell us? *Trends Ecol Evol* **20**: 435–440.

Beaumont MA., Balding DJ (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**: 969–980.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London B* **263**: 1619–1626.

Begg GA, Friedland KD, Pearce JB (1999). Stock identification and its role in stock assessment and fisheries management: an overview. *Fish Res* **43**: 1–8.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* **57**: 289–300.

Bert TM (1986). Speciation in western Atlantic stone crabs (genus *Menippe*): the role of geological processes and climatic events in the formation and distribution of species. *Mar Biol* **93**: 157–170.

Bohlmeyer DA, Gold JR (1991). Genetic studies in marine fishes II. A protein electrophoretic analysis of population structure in the red drum *Sciaenops ocellatus*. *Mar Biol* **108**: 197–206.

Borcard D, Gillet F, Legendre P (2011). *Numerical Ecology with R*. Springer: New York, NY.

Bourret V, Kent M, Primmer C (2013). SNP array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Mol Ecol* **22**: 532–551.

Bowen BW, Avise JC (1990). Genetic structure of Atlantic and Gulf of Mexico populations of sea bass, menhaden, and sturgeon: influence of zoogeographic factors and life-history patterns. *Mar Biol* **107**: 371–381.

Bradbury IR, Hubert S, Higgins B, Borza T, Bowman S, Paterson IG, *et al.* (2010). Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proc Biol Sci* **277**: 3725–34.

Bradbury IR, Hubert S, Higgins B, Bowman S, Borza T, Paterson IG, *et al.* (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evol Appl* **6**: 450–61.

Bricker, S., B. Longstaff, W. Dennison, A. Jones, K. Boicourt, *et al.* (2007) National estuarine eutrophication assessment database. URL: *http://ian.umces.edu/neea/*.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Carson EW, Bumguardner BW, Fisher M, Saillant E, Gold JR (2014). Spatial and temporal variation in recovery of hatchery-released red drum (*Sciaenops ocellatus*) in stock-enhancement of Texas bays and estuaries. *Fish Res* **151**: 191–198.

Carson EW, Karlsson S, Saillant E, Gold JR (2009). Genetic studies of hatchery-supplemented populations of red drum in four Texas bays. *North Am J Fish Manag* **29**: 1502–1510.

Carvalho G, Hauser L, Park S, Kingdom U (1994). Molecular genetics and the stock concept in fisheries (GR Carvalho and TJ Pitcher, Eds.). *Rev Fish Biol Fish* **4**: 326–350.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3 (Bethesda)* **1**: 171–82.

Cnaani A, Hallerman E, Ron M, Weller J (2003). Detection of a chromosomal region with two quantitative trait loci, affecting cold tolerance and fish size, in an F2 tilapia hybrid. *Aquaculture* **223**: 117–128.

Cnaani A, Zilberman N, Tinman S, Hulata G, Ron M (2004). Genome-scan analysis for quantitative trait loci in an F2 tilapia hybrid. *Mol Genet Genomics* **272**: 162–72.

Conover DO, Present TMC (1990). Countergradient variation in growth rate: compensation for length of the growing season among Atlantic silversides from different latitudes. *Oecologia* **83**: 316–324.

Corbin LJ, Blott SC, Swinburne JE, Vaudin M, Bishop SC, Woolliams JA (2010). Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Anim Genet* **41 Suppl 2**: 8–15.

Corbin LJ, Liu AYH, Bishop SC, Woolliams JA (2012). Estimation of historical effective population size using linkage disequilibria with marker data. *J Anim Breed Genet* **129**: 257–70.

Dahlberg MD (1970). Atlantic and Gulf of Mexico menhadens, genus Brevoortia (Pisces: Clupeidae). *Bull Florida State Museum* **15**: 91–162.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–8.

Danzmann RG, Gharbi K (2007). Linkage mapping in aquaculture species. In: Liu Z (ed) *Aquaculture Genome Technologies*, Blackwell Publishing Ltd: Oxford, UK, p 551.

Davey JW, Davey JL, Blaxter ML, Blaxter MW (2010). RADSeq: next-generation population genetics. *Brief Funct Genomics* **9**: 416–23.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

Davis JT (1990). *Red Drum Biology and Life History*. Southern Regional Aquaculture Center Publication No. 320. Texas A&M University. College Station, TX.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Mol Ecol Resour* **14**: 209–214.

Emerson S, Hedges J (2008). *Chemical Oceanography and the Marine Carbon Cycle*. Cambridge University Press: New York, NY.

England PR, Cornuet J-M, Berthier P, Tallmon DA, Luikart G (2006). Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv Genet* **7**: 303–308.

Estoup A, Largiader CR, Perrot E, Chourrout D (1996). Rapid one-tube DNA extraction for reliable PCR detection of fish polymorphic markers and transgenes. *Mol Mar Biol Biotechnol* **5**: 295–298.

Excoffier L, Hofer T, Foll M (2009). Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* **103**: 285–298.

Excoffier L, Lischer H (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**: 564–567.

Falconer DS, Mackay TFC (1996). *Introduction to Quantitative Genetics*, 4th edn. Pearson Education Limited. Harlow, England.

FAO (2015). Global aquaculture production: 1950-2014. URL: *http://www.fao.org/fishery/statistics/global-aquaculture-production/query/en*.

Flury C, Tapio M, Sonstegard T, Drögemüller C, Leeb T, Simianer H, *et al.* (2010). Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *J Anim Breed Genet* **127**: 339–47.

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

Fraser DJ, Bernatchez L (2001). Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Mol Ecol* **10**: 2741–2752.

Galeano CH, Fernandez AC, Franco-Herrera N, Cichy KA, McClean PE, Vanderleyden J, *et al.* (2011). Saturation of an intra-gene pool linkage map: towards a unified consensus linkage map for fine mapping and synteny analysis in common bean. *PLoS One* **6**: e28135.

Gjedrem T, Robinson N, Rye M (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: a review. *Aquaculture* **350-353**: 117–129.

GMFMC {Gulf of Mexico Fisheries Management Council} (1996). *Commercial Fishing Regulations for Gulf of Mexico Federal Waters*. GMFMC. Tampa, FL.

Gold JR, Burridge CP, Turner TF (2001). A modified stepping-stone model of population structure in red drum, *Sciaenops ocellatus* (Sciaenidae), from the northern Gulf of Mexico. *Genetica* **111**: 305–17.

Gold JR, Kedzie KM, Bohymeyer D, Jenkin JD, Karel WJ, Iida N, *et al.* (1988). Studies on the basic structure of the red drum (*Sciaenops ocellatus*) genome. *Contrib Mar Sci* **30**: 57–62.

Gold JR, Ma L, Saillant E, Silva PS, Vega RR (2008). Genetic effective size in populations of hatchery-raised red drum released for stock enhancement. *Trans Am Fish Soc* **137**: 1327–1334.

Gold JR, Richardson LR (1991). Genetic studies in marine fishes IV. An analysis (*Sciaenops ocellatus*) using mitochondrial DNA. *Fish Res* **12**: 213–241.

Gold JR, Richardson LR (1998). Population structure in greater amberjack, *Seriola dumerili*, from the Gulf of Mexico and the western Atlantic Ocean. *Fish Bull* **96**: 767–778.

Gold JR, Richardson LR, Furman C, King T (1993). Mitochondrial DNA differentiation and population structure in red drum (*Sciaenops ocellatus*) from the Gulf of Mexico and Atlantic Ocean. *Mar Biol* **185**: 175–185.

Gold JR, Richardson LR, Turner TF (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. *Mar Biol* **133**: 593–602.

Gold JR, Turner T (2002). Population structure of red drum (*Sciaenops ocellatus*) in the northern Gulf of Mexico, as inferred from variation in nuclear-encoded microsatellites. *Mar Biol* **140**: 249–265.

Gold JR, Pak E, DeVries DA (2002). Population structure of king mackerel (*Scomberomorus cavalla*) around peninsular Florida, as revealed by microsatellite DNA. *Fish Bull* **100**: 491–509.

Gold JR, Saillant E, Ebelt ND, Lem S (2009). Conservation genetics of gray snapper (*Lutjanus griseus*) in U.S. Waters of the northern Gulf of Mexico and western Atlantic Ocean. *Copeia* **2009**: 277–286.

Goodyear PG (1991). *Status of red drum stocks in the Gulf of Mexico*. National Marine Fisheries Service, Southeast Fisheries Center, Coastal Resources Division Contribution MIA-90/91-87. Miami, FL.

Gruenthal KM, Witting DA, Ford T, Neuman MJ, Williams JP, Pondella DJ, *et al.* (2014). Development and application of genomic tools to the restoration of green abalone in southern California. *Conserv Genet* **15**: 109–121.

Hauser L, Carvalho GR (2008). Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish Fish* **9**: 333–362.

Hayes BJ, Visscher PM, Mcpartlan HC, Goddard ME (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* **13**: 635–643.

Hemmer-Hansen J, Nielsen EE, Therkildsen NO, Taylor MI, Ogden R, Geffen AJ, *et al.* (2013). A genomic island linked to ecotype divergence in Atlantic cod. *Mol Ecol* **22**: 2653–2667.

Hermisson J (2009). Who believes in whole-genome scans for selection? *Heredity (Edinb)* **103**: 283–284.

Herrero-Medrano JM, Megens H-J, Groenen MAM, Ramis G, Bosse M, Pérez-Enciso M, *et al.* (2013). Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC Genet* **14**: 106.

Hill WG (1981). Estimation of effective population size from data on linkage disequilibrium. *Genet Res* **38**: 209–216.

Hindar K, Ryman N, Utter F (1991). Genetic effects of cultured fish on natural fish populations. *Can J Fish Aquat Sci* **48**: 945–957.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**: e1000862.

Hohenlohe PA, Phillips PC, Cresko WA (2010). Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *Int J Plant Sci* **171**: 1059–1071.

Hollenbeck CM, Portnoy DS, Gold JR (2012). Use of comparative genomics to develop EST-SSRs for red drum (*Sciaenops ocellatus*). *Mar Biotechnol* **14**: 672–80.

Hollenbeck CM, Portnoy DS, Gold JR (2015). A genetic linkage map of red drum (*Sciaenops ocellatus*) and comparison of chromosomal syntenies with four other fish species. *Aquaculture* **435**: 265–274.

Holt J, Godbout R, Arnold CR (1981). Effects of temperature and salinity on egg hatching and larval survival of red drum, *Sciaenops ocellata*. *Fish Bull* **73**: 569–573.

Howarth RW (1988). Nutrient limitation of net primary production in marine ecosystems. *Annu Rev Ecol Syst* **19**: 89–110.

Huang W, Xu B, Chan-Hilton A (2004). Forecasting flows in Apalachicola River using neural networks. *Hydrol Process* **18**: 2545–2564.

Huh OK, Wiseman WJ, Rouse LJ (1981). Intrusion of Loop Current waters onto the west Florida continental shelf. *J Geophys Res Ocean* **86**: 4186–4192.

Hulata G (2001). Genetic manipulations in aquaculture: a review of stock improvement by classical and modern technologies. *Genetica* **111**: 155–173.

International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**: 789–796.

Johnson DR, Perry HM, Lyczkowski-Shultz J, Hanisko D (2009). Red snapper larval transport in the northern Gulf of Mexico. *Trans Am Fish Soc* **138**: 458–470.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–5.

Karlsson S, Ma L, Saillant E, Gold J (2007). Tests of Mendelian segregation and linkage-group relationships among 31 microsatellite loci in red drum, *Sciaenops ocellatus*. *Aquac Int* **15**: 383–391.

Karlsson S, Saillant E, Bumguardner BW, Vega RR, Gold JR (2008). Genetic identification of hatchery-released red drum in Texas bays and estuaries. *North Am J Fish Manag* **28**: 1294–1304.

Karlsson S, Saillant E, Gold JR (2009). Population structure and genetic variation of lane snapper (*Lutjanus synagris*) in the northern Gulf of Mexico. *Mar Biol* **156**: 1841–1855.

Khan MA, Han Y, Zhao YF, Troggio M, Korban SS (2012). A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. *PLoS One* **7**: e47864.

Krzywinski M, Schein J, Birol İ (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.

Kucuktas H, Wang S, Li P, He C, Xu P, Sha Z, *et al.* (2009). Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. *Genetics* **181**: 1649–60.

Larson WA, Seeb LW, Everett M V., Waples RK, Templin WD, Seeb JE (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evol Appl* **7**: 355–369.

Leder E, Danzmann R, Ferguson M (2006). The candidate gene, Clock, localizes to a strong spawning time quantitative trait locus region in rainbow trout. *J Hered* **97**: 74–80.

Lee B-YB, Hulata G, Kocher TD (2004). Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*). *Heredity (Edinb)* **92**: 543–9.

Lee B, Penman D, Kocher T (2003). Identification of a sex-determining region in Nile tilapia (*Oreochromis niloticus*) using bulked segregant analysis. *Anim Genet* **34**: 379–383.

Leidig JM, Shervette VR, McDonough CJ, Darden TL (2015). Genetic population structure of black drum in US waters. *North Am J Fish Manag* **35**: 464–477.

Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**: 175–195.

Li C, Wang R, Su B, Luo Y, Terhune J, Beck B, *et al.* (2013). Evasion of mucosal defenses during *Aeromonas hydrophila* infection of channel catfish (*Ictalurus punctatus*) skin. *Dev Comp Immunol* **39**: 447–55.

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

Lotterhos KE, Whitlock MC (2014). Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Mol Ecol* **23**: 2178–2192.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003). The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* **4**: 981–94.

Luikart G, Ryman N, Tallmon DA., Schwartz MK, Allendorf FW (2010). Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conserv Genet* **11**: 355–373.

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits*, 1st edn. Sinauer Associates: Sunderland, MA.

Matlock GC (1987). The life history of the red drum. In: Chamberlain G, Miget R, Haby M (eds) *Manual of Red Drum Aquaculture*, Texas Agricultural Extension Service and Sea Grant College Program, Texas A&M University: College Station, TX, pp 1–47.

Matlock GC (1990). Preliminary results of red drum stocking in Texas. In: *Marine Farming and Enhancement: Proceedings of the 15th US Japan Meeting on Aquaculture. NOAA Technical Report NMFS*, Vol 85, pp 11–15.

McClure MR, McEachran JD (1992). Hybridization between *Prionotus alatus* and *P. paralatus* in the northern Gulf of Mexico (Pisces: Triglidae). *Copeia*: 1039–1046.

McEachron LW, McCarty CE, Vega RR (1995). Beneficial uses of marine fish hatcheries: enhancement of red drum in Texas coastal waters. In: *American Fisheries Society Symposium*, Vol 15, pp 161–166.

McKinney G, Seeb L, Larson W, Gomez-Uchida D, Limborg M, Brieuc M, *et al.* (2015). An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (*Oncorhynchus tshawytscha*). *Mol Ecol Resour*: doi: 10.1111/1755-0998.12479

Meirmans PG (2015). Seven common mistakes in population genetics and how to avoid them. *Mol Ecol* **24**: 3223–3231.

Michaelsen SA (2015). Assessing the spatio-temporal genetic stock structure of red drum (*Sciaenops ocellatus*) in the northern Gulf of Mexico (*Thesis*). Southeastern Louisiana University.

Moen T, Agresti J, Cnaani A (2004). A genome scan of a four-way tilapia cross supports the existence of a quantitative trait locus for cold tolerance on linkage group 23. *Aquac Res* **35**: 893–904.

Murrell P (2005). *R graphics*, 1st edn. CRC Press: Boca Raton, FL.

Narum SR, Hess JE (2011). Comparison of F_{ST} outlier tests for SNP loci under selection. *Mol Ecol Resour* **11 Suppl 1**: 184–94.

Navarro A, Barton NH (2003). Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution* **57**: 447–459.

Nei M, Tajima F (1981). Genetic drift and estimation of effective population size. *Genetics* **98**: 625–640.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009). Population genomics of marine fishes: identifying adaptive variation in space and time. *Mol Ecol* **18**: 3128–3150.

NMFS {National Marine Fisheries Service} (2015a). Recreational fisheries statistics queries. URL: *http://www.st.nmfs.noaa.gov/st1/recreational/queries/*.

NMFS {National Marine Fisheries Service} (2015b). *Fisheries of the United States 2014: Current Fishery Statistics No. 2014*. National Oceanographic and Atmospheric Association Silver Springs, MD.

Nosil P, Egan SP, Funk DJ (2008). Heterogeneous genomic differentiation between walking-stick ecotypes: ‘isolation by adaptation’ and multiple roles for divergent selection. *Evolution (NY)* **62**: 316–336.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009). Divergent selection and heterogeneous genomic divergence. *Mol Ecol* **18**: 375–402.

NREL {National Renewable Energy Laboratory} (2015). Wind maps. URL: *http://www.nrel.gov/gis/wind.html*.

O'Brien S (1991). Mammalian genome mapping: lessons and prospects. *Curr Opin Genet Dev* **1**: 105–111.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, *et al.* (2015). vegan: community ecology package. R package version 2.2-1. URL: *http://CRANR-project.org/package=vegan*.

Orsini L, Mergeay J, Vanoverbeke J, De Meester L (2013). The role of selection in driving landscape genomic structure of the waterflea *Daphnia magna*. *Mol Ecol* **22**: 583–601.

Pacitti D, Wang T, Martin SAM, Sweetman J, Secombes CJ (2014). Insights into the fish thioredoxin system: expression profile of thioredoxin and thioredoxin reductase in rainbow trout (*Oncorhynchus mykiss*) during infection and in vitro stimulation. *Dev Comp Immunol* **42**: 261–77.

Paradis E (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**: 419–420.

Pattillo M, Czapla T, Nelson D, Monaco M (1997). *Distribution and Abundance of Fishes and Invertebrates in Gulf of Mexico Estuaries, Volume II: Species Life History Summaries*. US Dept. of Commerce, National Oceanic and Atmospheric Administration, National Ocean Service, Office of Ocean Resources Conservation and Assessment, Strategic Environmental Assessments Division, Biogeographic Characterization Branch. Rockville, MD.

Paytan A, McLaughlin K (2007). The oceanic phosphorus cycle. *Chem Rev* **107**: 563–576.

Peng B, Kimmel M (2005). simuPOP: A forward-time population genetics simulation environment. *Bioinformatics* **21**: 3686–3687.

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.

Poh YP, Domingues VS, Hoekstra HE, Jensen JD (2014). On the prospect of identifying adaptive loci in recently bottlenecked populations. *PLoS One* **9**: 1–9.

Pollak E (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**: 531–548.

Portnoy DS, Gold JR (2012). Evidence of multiple vicariance in a marine suture-zone in the Gulf of Mexico. *J Biogeogr* **39**: 1499–1507.

Portnoy D, Hollenbeck C, Belcher C, Driggers III W, Frazier B, Gelsleichter J, *et al.* (2014). Contemporary population structure and post-glacial genetic demography in a migratory marine species, the blacknose shark, *Carcharhinus acronotus*. *Mol Ecol* **23**: 5480–5495.

Portnoy DS, Hollenbeck CM, Renshaw M a., Gold JR (2011). Microsatellite panels for gene localization in red drum, *Sciaenops ocellatus*. *Aquaculture* **319**: 505–508.

Portnoy DS, Puritz JB, Hollenbeck CM, Gelsleichter J, Chapman D, Gold JR (2015). Selection and sex-biased dispersal in a coastal shark: the influence of philopatry on adaptive variation. *Mol Ecol* **24**: 5877–5885.

Portnoy DS, Renshaw MA, Hollenbeck CM, Gold JR (2010). A genetic linkage map of red drum, *Sciaenops ocellatus*. *Anim Genet* **41**: 630–641.

Poulin R, Morand S (2000). The diversity of parasites. *Q Rev Biol*: 277–293.

Puritz JB, Hollenbeck CM, Gold JR (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**: e431.

R Core Team (2015). R: a language and environment for statistical computing. *R Found Stat Comput* **1**: 409.

Rabalais N, Turner R, JustiĆ D, Dortch Q, Wiseman W, Sen Gupta B (1996). Nutrient changes in the Mississippi River and system responses on the adjacent continental shelf. *Estuaries and Coasts* **19**: 386–407.

Ramsey PR, Wakeman JM (1987). Population structure of *Sciaenops ocellatus* and *Cynoscion nebulosus* (Pisces: Sciaenidae): biochemical variation, genetic subdivision and dispersal. *Copeia* **1987**: 682–695.

Reeb C a., Avise JC (1990). A genetic discontinuity in a continuously distributed species: Mitochondrial DNA in the American oyster, *Crassostrea virginica*. *Genetics* **124**: 397–406.

Renshaw M, Hollenbeck C, Gold JR (2012). Isolation of microsatellite markers from red drum, *Sciaenops ocellatus*, and characterization in red drum and spotted seatrout, *Cynoscion nebulosus*. *Mol Ecol Resour* **12**: 570–572.

Roesti M, Salzburger W, Berner D (2012). Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol Biol* **12**: 94.

Rooker JR, Holt SA (1997). Utilization of subtropical seagrass meadows by newly settled red drum *Sciaenops ocellatus*: patterns of distribution and growth. *Mar Ecol Prog Ser* **158**: 139–149.

Ross JL, Stevens TM, Vaughan DS (1995). Age, growth, mortality, and reproductive biology of red drums in North Carolina waters. *Trans Am Fish Soc* **124**: 37–54.

Ruzzante DE, Taggart CT, Cook D (1999). A review of the evidence for genetic structure of cod (*Gadus morhua*) populations in the NW Atlantic and population affinities of larval cod off Newfoundland and the Gulf of St. Lawrence. *Fish Res* **43**: 79–97.

Ryther JH (1969). Photosynthesis and fish production in the sea. *Science* **166**: 72–76.

Sarropoulou E, Franch R, Louro B, Power DM, Bargelloni L, Magoulas A, *et al.* (2007). A gene-based radiation hybrid map of the gilthead sea bream *Sparus aurata* refines and exploits conserved synteny with *Tetraodon nigroviridis*. *BMC Genomics* **8**: 44.

Scharf FS (2000). Patterns in abundance, growth, and mortality of juvenile red drum across estuaries on the Texas coast with implications for recruitment and stock enhancement. *Trans Am Fish Soc* **129**: 1207–1222.

Schulte PM (2007). Responses to environmental stressors in an estuarine fish: interacting stressors and the impacts of local adaptation. *J Therm Biol* **32**: 152–161.

Schwartz MK, Luikart G, Waples RS (2007). Genetic monitoring as a promising tool for conservation and management. *Trends Ecol Evol* **22**: 25–33.

Seyoum S, Tringali MD, Bert TM, McElroy D, Stokes R (2000). An analysis of genetic population structure in red drum, *Sciaenops ocellatus*, based on mtDNA control region sequences. *Fish Bull* **98**: 127–138.

Shirak A, Seroussi E, Cnaani A, Howe AE, Domokhovsky R, Zilberman N, *et al.* (2006). Amh and Dmrta2 genes map to tilapia (*Oreochromis* spp.) linkage group 23 within quantitative trait locus regions for sex determination. *Genetics* **174**: 1573–81.

Smith TIJ, Jenkins WE, Denson MR, Collins MR (2001). Stock enhancement research with anadromous and marine fishes in South Carolina. *Ecol Aquac species Enhanc Stock*: 175.

Sonesson A (2007). Possibilities for marker-assisted selection in aquaculture breeding schemes. In: Guimarães EP, Ruane J, Scherf B, Sonnino A, Dargie J (eds) *Marker-Assisted Selection: Current Status and Future Perspectives in Crops, Livestock, Forestry, and Fish*, FAO. Rome.

Southwick Associates (2013). *Sportfishing in America: An Economic Force for Conservation*. Alexandria, VA.

Storz JF (2005). Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol* **14**: 671–688.

Sved JA (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* **2**: 125–141.

Sved JA, Cameron EC, Gilchrist AS (2013). Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. *PLoS One* **8**: e69078.

Tao WJ, Boulding EG (2003). Associations between single nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (*Salvelinus alpinus* L.). *Heredity (Edinb)* **91**: 60–9.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, *et al.* (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**: 520–6.

Teshima KM, Coop G, Przeworski M (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**: 702–712.

Tringali MD, Leber KM, Halstead WG, Mcmichael R, O'Hop J, Winner B, *et al.* (2008). Marine stock enhancement in Florida: a multi-disciplinary, stakeholder-supported, accountability-based approach. *Rev Fish Sci* **16**: 51–57.

Turner TL, Hahn MW, Nuzhdin S V (2005). Genomic islands of speciation in *Anopheles gambiae*. *PLOS Biol* **3**: e285.

Uchupi E (1975). Physiography of the Gulf of Mexico and Caribbean Sea. In: Nairn, Alan (ed) *The Ocean Basins and Margins, Volume 3: The Gulf of Mexico and the Caribbean*, Springer: New York, NY.

USDA (2014). *Census of Aquaculture - 2013*. United States Department of Agriculture. Washington, DC.

van Ooijen JW (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet Res (Camb)* **93**: 343–9.

Vangestel C, Mergeay J, Dawson DA, Callens T, Vandomme V, Lens L (2012). Genetic diversity and population structure in contemporary house sparrow populations along an urbanization gradient. *Heredity (Edinb)* **109**: 163–172.

Vega RR, Chavez C, Stolte CJ, Abrego D (2003). *Marine fish distribution report, 1991-1999*. Texas Parks and Wildlife. Austin, TX.

Waples RS (1995). Evolutionarily Significant Units and the Conservation of Biological Diversity under the Endangered Species Act. *Am Fish Soc Symp* **17**: 8–27.

Waples RS (1998). Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *J Hered* **89**: 438–450.

Waples RS (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv Genet* **7**: 167–184.

Waples RS, Do C (2008). LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Resour* **8**: 753–6.

Waples RS, Do C (2010). Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evol Appl* **3**: 244–262.

Waples RK, Larson WA, Waples RS (2015). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity (Edinb), in press*.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution (NY)* **38**: 1358–1370.

Weir BS, Hill WG (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**: 477–488.

Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York, NY.

Wilhelm O, Ewing M (1972). Geology and history of the Gulf of Mexico. *Geol Soc Am Bull* **83**: 575–600.

Wiseman Jr WJ, Dinnel SP (1988). Shelf currents near the mouth of the Mississippi River. *J Phys Oceanogr* **18**: 1287–1291.

Woodward AG (2000). Red drum stock enhancement in Georgia: a responsible approach. Coastal Resources Division, Georgia Department of Natural Resources. Brunswick, GA.

Wright S (1943). Isolation by distance. *Genetics* **28**: 114–138.

Wu CI (2001). The genic view of the process of speciation. *J Evol Biol* **14**: 851–865.

Wu Y, Close TJ, Lonardi S (2011). Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Trans Comput Biol Bioinform* **8**: 381–94.

Wu C, Zhang D, Kan M, Lv Z, Zhu A, Su Y, *et al.* (2014). The draft genome of the large yellow croaker reveals well-developed innate immunity. *Nat Commun* **5**: 5227.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Others (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**: 565–569.

Zhang M, Sun L (2011). The tissue factor pathway inhibitor 1 of *Sciaenops ocellatus* possesses antimicrobial activity and is involved in the immune response against bacterial infection. *Dev Comp Immunol* **35**: 247–252.

Zhao L, Hu Y-H, Sun J-S, Sun L (2011). The high mobility group box 1 protein of *Sciaenops ocellatus* is a secreted cytokine that stimulates macrophage activation. *Dev Comp Immunol* **35**: 1052–8.

Zhou L, Bai R, Tian J, Liu X, Lu D, Zhu P, *et al.* (2009). Bioinformatic comparisons and tissue expression of the neuronal nitric oxide synthase (nNOS) gene from the red drum (*Sciaenops ocellatus*). *Fish Shellfish Immunol* **27**: 577–84.

APPENDIX A

SUPPLEMENTAL TABLES

Supplemental Table A1 Pairwise matrix of approximate coastline geographic distance. Measurements were calculated with Google Earth® and are reported in in kilometers (km).

	LLM	MAT	SAB	MIS	APA	CEK	CHA	IND	HAR	WAS	SCA
LLM	--										
MAT	290	--									
SAB	567	277	--								
MIS	1260	970	693	--							
APA	1687	1397	1120	427	--						
CEK	1961	1671	1394	701	274	--					
CHA	2248	1958	1681	988	561	287	--				
IND	2985	2695	2418	1725	1298	1024	737	--			
HAR	3306	3016	2739	2046	1619	1345	1058	321	--		
WAS	3391	3101	2824	2131	1704	1430	1143	406	85	--	
SCA	3537	3247	2970	2277	1850	1576	1289	552	231	146	--

Supplemental Table A2 Environmental variables for each sampling locality, obtained from the National Estuarine Eutrophication Assessment database. The descriptions of each variable can be found in the metadata from the database.

Variable	LLM	MAT	SAB	MIS	APA	CEK	CHA	IND	HAR	WAS	SCA
Estuary Area (km2)	1308	1115	265	1581	593	165	502	866	39	88	85
Tidal Fresh Zone Area (km2)	0	2	8	0	46	7	1	0	5	11	1
Mixing Zone Area (km2)	27	918	257	1409	273	110	201	109	29	38	58
Saltwater Zone Area (km2)	1281	195	0	172	274	48	300	757	5	38	25
Estuary Depth (m)	0.76	1.41	2.49	2.43	1.81	1.17	1.63	0.77	1.96	3.35	4.99
Estuary Perimeter (km)	1413	790	355	553	402	345	567	1383	215	320	313
Percent Estuary Open (%)	0.04	0.76	0.17	1.99	2.99	11.89	2.64	0.21	1.63	1.75	0.83
Catchment Area (km2)	13165	121762	53674	4050	52214	25989	8134	3093	36962	12133	41116
Catchment Mean Elev. (m)	43	619	86	33	148	51	22	5	133	70	216
Catchment Max Elev. (m)	257	1374	236	124	1250	139	75	14	459	244	1679
Catchment/Estuary Area Ratio	10.1	109.2	202.5	2.6	88.1	157.5	16.2	3.6	947.7	137.9	483.7
Total Land Cover (km2)	13048.362	121234.778	51999.1897	39069.94	51483.7999	25837.6779	11447.736	2944.811	36754.50998	12058.9776	39857.31
Population (#)	616541	1432800	1230500	215299	2738086	417564	397072	471807	1681584	268166	3139518
Pop / Estuary Area (#.km-2)	471.4	1285	4643.4	136.2	4617.3	2530.7	791	544.8	43117.5	3047.3	36935.5
Tide Height (m)	0.4	0.2	0.47	0.51	0.58	0.76	0.65	0.32	1.9	1.93	1.45
Tide Ratio	0.53	0.14	0.19	0.21	0.32	0.65	0.4	0.42	0.97	0.58	0.29
Stratification Ratio	0.00146	0.04799	0.36975	0.00793	0.0976	0.10669	0.0085	0.00683	0.23917	0.02307	0.05799
Percent Freshwater (%)	0	0.2	3.1	0	7.8	4.1	0.2	0	12	12.6	1.4
Percent Mixed Water (%)	2.1	82.3	96.9	89.1	46.1	66.6	40.1	12.6	75.1	43.7	68.8
Percent Seawater (%)	97.9	17.5	0	10.9	46.2	29.3	59.7	87.4	13	43.7	29.8
Average Salinity (psu)	29	16	12	15	19	17	22	29	13	18	17
Tidal Exchange (days)	233	82	10	336	8	4	59	36	1	19	16
Tidal Freshwater Flush (d)	4	38	10	23	4	1	3	3	0	1	5
Daily FW/Est Area (m.d-1)	1.575	8.906	166.415	1.961	107.926	155.758	10.1	3.441	876.923	84.886	161.176
Daily Freshwater (m3.d-1)	2060000	9930000	44100000	3100000	64000000	25700000	5070000	2980000	34200000	7470000	13700000
Flow / Estuary Area (m.d-1)	1.575	8.906	166.415	1.961	107.926	155.758	10.1	3.441	876.923	84.886	161.176
Total FW Volume (1.d-1)	0.00074	0.00658	0.06743	0.00161	0.06044	0.13392	0.00655	0.00549	0.44801	0.02569	0.03256
Daily Precipitation (m3.d-1)	2.41E+06	3.27E+06	1.05E+06	6.88E+06	2.30E+06	571000	1.72E+06	3.06E+06	138000	306000	293000
Daily Evaporation (m3.d-1)	3.73E+06	2.86E+06	655000	3.80E+06	1.43E+06	417000	1.43E+06	2.38E+06	91800	203000	182000
Daily Precip / Est Area (mm.d-1)	1.843	2.933	3.962	4.352	3.879	3.461	3.426	3.533	3.538	3.477	3.447
Daily Evap / Est Area (mm.d-1)	2.852	2.565	2.472	2.404	2.411	2.527	2.849	2.748	2.354	2.307	2.141
Flow (m3.d-1)	2.06E+06	1.26E+07	5.97E+07	1.84E+06	6.56E+07	3.37E+07	2.46E+06	1.22E+06	3.88E+07	1.17E+07	3.60E+07
Air Temp Mean (C)	23.3	21.3	20.6	20.2	20.8	21.8	23.2	22.6	20.4	19.7	18.6
Air Temp Std Dev (C)	4.8	5.9	6.2	6.2	5.7	5	3.8	4	5.8	6.3	6.4
Frost Days (#)	3	9	15	21	16	13	1	3	28	29	41

Supplemental Table A2 continued

Variable	LLM	MAT	SAB	MIS	APA	CEK	CHA	IND	HAR	WAS	SCA
Wind Speed (m.sec-1)	7.1	7	6.4	6.2	6	6	5.9	6.5	6.7	6.8	7.1
Sea Surface Temp Mean (C)	24.6	24.1	23.4	23.2	23.7	24	25.5	26.1	23.8	23.8	23.9
Sea Surface Temp Std Dev (C)	3.5	4.1	4.8	4.8	4.2	4	3.1	2.4	3.4	3.3	3.1
Ocean Salinity Mean (psu)	35.1	34.6	33.9	32.7	35.3	35.2	35.9	36.1	35.2	35.2	35.6
Ocean Salinity Max (psu)	36.6	36.2	35.8	35.1	36.4	36	36.3	36.3	35.8	35.8	36
Ocean Salinity Min (psu)	31.9	30.5	30.7	30.2	32.9	33.3	34.7	36	34.6	34.5	34.9
Oceanic DIP (µM)	0.2	0.21	0.32	0.27	0.13	0.12	0.1	0.16	0.17	0.16	0.16
Oceanic NO3 (µM)	0.12	0.14	0.7	1.95	0.87	0.59	0.58	0.27	0.29	0.29	0.33
TSS (tonne.y-1)	3.18E+06	1.14E+06	811000	563000	127000	175000	140000	39700	1960	17400	32200
TN (kg.y-1)	9.13E+06	9.13E+06	2.34E+07	1.62E+06	2.59E+07	5.78E+06	1.85E+06	1.61E+06	1.53E+07	6.40E+06	2.00E+06
TP (kg.y-1)	337900	830800	1.36E+06	166780	970300	2.46E+06	290780	187860	799800	440200	278380
TSS/Est Area (tonne.km-2.y-1)	2431.2	1022.4	3060.4	356.1	214.2	1060.6	278.9	45.8	50.3	197.7	378.8
TN/Est Area (kg.km-2.y-1)	6978.6	8186.5	88226.4	1027.2	43676.2	35042.4	3681.3	1859.1	391282.1	72704.5	23552.9
TP/Est Area (kg.km-2.y-1)	258.33	745.11	5135.47	105.49	1636.26	14917.58	579.24	216.93	20507.69	5002.27	3275.06

Supplemental Table A3 A summary of data-filtering procedures for the population genomics dataset; rows refer to each filtering step and columns refer to statistics for each step. In column names, 'sites' refer to individual polymorphisms (SNPs, indels, or complex polymorphisms), 'loci' refers to RAD contigs (each of which may contain multiple sites), and 'Inds' refers to individuals. 'Start', 'End', and 'Removed' refer, respectively, to the number of each unit before the filtering step, the number after the filtering step, and the number removed with the filter.

Filter	Start sites	End sites	Removed sites	Start loci	End loci	Removed loci	Start Inds	End Inds	Removed Inds
Genotype depth < 10	430466	430466	0	33170	33170	0	568	568	0
Mean site quality < 20	430466	361415	69051	33170	32988	182	568	568	0
Mean site call rate < 0.5	361415	98309	263106	32988	8647	24341	568	568	0
Ind depth < 10 & call rate < 0.25	98309	98309	0	8647	8647	0	568	551	17
Mean site call rate < 0.75	98309	69484	28825	8647	5946	2701	551	551	0
Ind call rate < 0.6	69484	69484	0	5946	5946	0	551	531	20
Minor allele frequency < 0.05	69484	7890	61594	5946	3751	2195	531	531	0
Discordant sites between duplicates	7890	7811	79	3751	3740	11	531	531	0
Remove duplicate individuals	7811	7811	0	3740	3740	0	531	526	5
Remove known hatchery individuals	7811	7811	0	3740	3740	0	526	504	22
Remove related individuals	7811	7811	0	3740	3740	0	504	496	8
dDocent_filters script	7811	7351	460	3740	3633	107	496	496	0
Decomposed to allelic primitives	7351	7839	-488	3633	3633	0	496	496	0
Hardy-Weinberg equilibrium	7839	7539	300	3633	3563	70	496	496	0
Mean site call rate < 0.75	7539	7539	0	3563	3563	0	496	470	26
Mean site call rate by locality < 0.85	7539	3689	3850	3563	1804	1759	470	470	0
Mean site call rate overall < 0.95	3689	3642	47	1804	1784	20	470	470	0
Haplotyping	3642	2874	768	1784	1543	241	470	470	0
Manual Inspection and Filtering	2874	2860	14	1543	1539	4	470	462	8

Supplemental Table A4 Estimates of pairwise F_{ST} for 11 sampled localities, using datasets consisting of all loci (n = 1,539), only neutral loci (n = 1,393), and only outlier loci (n = 146). Estimates of pairwise F_{ST} (lower diagonal); Probability (P) that F_{ST} = 0 (upper diagonal). Significance was assessed by permutation test with 10,000 permutations. Estimates in bold represent significant values following correction, using a false discovery rate (FDR) of 0.05.

All Loci

	LLM	MAT	SAB	MIS	APA	CEK	CHA	IND	HAR	WAS	SCA
LLM	--	0.4825	0.8327	0.3551	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MAT	0.0005	--	0.7704	0.3161	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SAB	-0.0001	0.0000	--	0.4289	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MIS	0.0006	0.0006	0.0002	--	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
APA	**0.0159**	**0.0149**	**0.0149**	**0.0125**	--	0.0782	0.0002	0.0000	0.0000	0.0000	0.0000
CEK	**0.0231**	**0.0222**	**0.0223**	**0.0193**	0.0011	--	0.5191	0.0000	0.0000	0.0000	0.0000
CHA	**0.0265**	**0.0251**	**0.0253**	**0.0227**	**0.0025**	0.0004	--	0.0000	0.0000	0.0000	0.0000
IND	**0.0152**	**0.0140**	**0.0136**	**0.0128**	**0.0068**	**0.0106**	**0.0123**	--	0.0000	0.0031	0.0049
HAR	**0.0131**	**0.0116**	**0.0129**	**0.0114**	**0.0091**	**0.0121**	**0.0142**	**0.0029**	--	0.0084	0.3804
WAS	**0.0141**	**0.0128**	**0.0133**	**0.0110**	**0.0086**	**0.0122**	**0.0136**	**0.0019**	**0.0020**	--	0.7662
SCA	**0.0130**	**0.0115**	**0.0119**	**0.0100**	**0.0075**	**0.0109**	**0.0128**	**0.0014**	0.0006	0.0001	--

Supplemental Table A4 continued

Neutral Loci

	LLM	MAT	SAB	MIS	APA	CEK	CHA	IND	HAR	WAS	SCA
LLM	--	0.7251	0.8731	0.8625	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MAT	0.0002	--	0.9355	0.8256	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SAB	-0.0002	-0.0004	--	0.5500	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MIS	-0.0002	-0.0001	0.0001	--	0.1349	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
APA	**0.0025**	**0.0023**	**0.0024**	0.0011	--	0.9224	0.6476	0.0000	0.0000	0.0000	0.0000
CEK	**0.0037**	**0.0039**	**0.0040**	**0.0027**	-0.0002	--	0.8543	0.0000	0.0000	0.0000	0.0000
CHA	**0.0042**	**0.0032**	**0.0037**	**0.0030**	0.0003	0.0000	--	0.0000	0.0000	0.0000	0.0000
IND	**0.0063**	**0.0060**	**0.0053**	**0.0058**	**0.0034**	**0.0051**	**0.0054**	--	0.0150	0.2711	0.1505
HAR	**0.0050**	**0.0043**	**0.0053**	**0.0045**	**0.0045**	**0.0044**	**0.0046**	**0.0019**	--	0.0818	0.7526
WAS	**0.0062**	**0.0056**	**0.0057**	**0.0042**	**0.0047**	**0.0048**	**0.0047**	0.0008	0.0014	--	0.8809
SCA	**0.0056**	**0.0049**	**0.0051**	**0.0041**	**0.0034**	**0.0040**	**0.0041**	0.0008	0.0002	-0.0001	--

Outlier Loci

	LLM	MAT	SAB	MIS	APA	CEK	CHA	IND	HAR	WAS	SCA
LLM	--	0.0256	0.3258	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MAT	**0.0034**	--	0.0117	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SAB	0.0012	**0.0041**	--	0.1308	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MIS	**0.0087**	**0.0074**	0.0020	--	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
APA	**0.1350**	**0.1295**	**0.1258**	**0.1122**	--	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CEK	**0.1846**	**0.1781**	**0.1753**	**0.1574**	**0.0141**	--	0.0053	0.0000	0.0000	0.0000	0.0000
CHA	**0.2100**	**0.2075**	**0.2024**	**0.1839**	**0.0234**	**0.0045**	--	0.0000	0.0000	0.0000	0.0000
IND	**0.0966**	**0.0893**	**0.0886**	**0.0755**	**0.0373**	**0.0591**	**0.0737**	--	0.0000	0.0000	0.0001
HAR	**0.0889**	**0.0819**	**0.0841**	**0.0737**	**0.0506**	**0.0803**	**0.0971**	**0.0122**	--	0.0017	0.0063
WAS	**0.0882**	**0.0823**	**0.0831**	**0.0724**	**0.0439**	**0.0781**	**0.0904**	**0.0120**	**0.0074**	--	0.1382
SCA	**0.0812**	**0.0733**	**0.0744**	**0.0635**	**0.0435**	**0.0718**	**0.0882**	**0.0068**	**0.0052**	0.0020	--

Supplemental Table A5 Summary of candidate genes for loci contributing primarily to PC1 and PC2 in the PCA of outlier loci. Gene and protein annotations are predictions from the genome of the large yellow croaker in the RefSeq genome assembly annotation (GenBank accession: ASM74293v1). The scaffold accession and start/stop coordinates refer to the same assembly version. The RAD locus for each entry is the nearest outlier locus in proximity to the gene.

Protein Name	RAD Locus	Scaffold	Start	Stop	Strand	Gene ID	Gene Name	Protein Accession	Length
PC1									
neurexin-2-like	Contig_29149	NW_011322530.1	163264	483252	+	104919663	LOC104919663	XP_010730030.1	1061
mitogen-activated protein kinase kinase kinase 11	Contig_29149	NW_011322530.1	528129	569008	-	104919676	map3k11	XP_010730045.1	1044
histone-lysine N-methyltransferase 2A	Contig_41916	NW_011322458.1	1398499	1434688	-	104936774	kmt2a	XP_010751167.1	4489
ubiquitin conjugation factor E4 A isoform X2	Contig_41916	NW_011322458.1	1442852	1451063	-	104936480	ube4a	XP_010750832.1	1077
bicaudal D-related protein 1-like	Contig_41916	NW_011322458.1	1459180	1474837	-	104936494	LOC104936494	XP_010750843.1	538
protein FAM98B-like	Contig_41916	NW_011322458.1	1480104	1484531	+	104936503	LOC104936503	XP_010750854.1	349
coiled-coil-helix-coiled-coil-helix domain-containing protein 2, mitochondrial	Contig_41916	NW_011322458.1	1486676	1488853	+	104936513	chchd2	XP_010750867.1	169
phosphorylase b kinase gamma catalytic chain, skeletal muscle/heart isoform	Contig_41916	NW_011322458.1	1492117	1496362	+	104936522	phkg1	XP_010750875.1	394
phosphoserine phosphatase isoform X1	Contig_41916	NW_011322458.1	1502505	1505300	+	104936532	psph	XP_010750888.1	242
protein NipSnap homolog 2	Contig_41916	NW_011322458.1	1506732	1512502	-	104936548	gbas	XP_010750907.1	286
elastin-like	Contig_41916	NW_011322458.1	1571187	1602279	+	104936783	LOC104936783	XP_010751178.1	472
elastin-like	Contig_41916	NW_011322458.1	1607229	1617573	+	104936793	LOC104936793	XP_010751189.1	254
intraflagellar transport protein 74 homolog	Contig_53440	NW_011323504.1	2451	12001	+	104934431	ift74	XP_010748393.1	604
zinc finger CCHC domain-containing protein 2	Contig_53440	NW_011323504.1	17169	39560	-	104934432	zcchc2	XP_010748394.1	1155
collagen alpha-1(XXI) chain-like	Contig_53440	NW_011323504.1	52247	111426	+	104934433	LOC104934433	XP_010748396.1	816
dermatan-sulfate epimerase-like protein	Contig_30010	NW_011323737.1	60418	64050	+	104936053	dsel	XP_010750322.1	1210
N-lysine methyltransferase setd6-like	Contig_30010	NW_011323737.1	65618	72312	+	104936054	LOC104936054	XP_010750323.1	330
CCR4-NOT transcription complex subunit 1 isoform X9	Contig_30010	NW_011323737.1	74834	91686	-	104936050	cnot1	XP_010750319.1	2372
actin cytoskeleton-regulatory complex protein PAN1-like	Contig_30010	NW_011323737.1	102129	114325	+	104936052	LOC104936052	XP_010750320.1	424
relaxin receptor 1	Contig_12765	NW_011323281.1	69440	137956	+	104932580	rxfp1	XP_010746150.1	790
protein diaphanous homolog 3	Contig_42920	NW_011322786.1	306926	462150	-	104926257	diaph3	XP_010738373.1	1206
tudor domain-containing protein 3 isoform X1	Contig_42920	NW_011322786.1	469379	487009	+	104926258	tdrd3	XP_010738374.1	788
catenin delta-2-like	Contig_66331	NW_011323039.1	43512	165020	-	104929872	LOC104929872	XP_010742805.1	1181
cAMP-dependent protein kinase inhibitor alpha	Contig_66331	NW_011323039.1	213707	214752	+	104929873	pkia	XP_010742806.1	79
tyrosine-protein kinase Lyn isoform X1	Contig_66331	NW_011323039.1	221150	229509	+	104929874	lyn	XP_010742808.1	511
uroporphyrinogen decarboxylase-like, partial	Contig_66331	NW_011323039.1	232289	234278	+	104929877	LOC104929877	XP_010742811.1	126
dynein heavy chain 10, axonemal, partial	Contig_67276	NW_011324625.1	537	18960	+	104939268	dnah10	XP_010754084.1	3255
kelch-like protein 20	Contig_33200	NW_011323582.1	151033	152396	-	104934996	LOC104934996	XP_010749079.1	117
CUB and sushi domain-containing protein 3	Contig_67449	NW_011322428.1	472682	719653	-	104919588	csmd3	XP_010729953.1	3552
proteasome assembly chaperone 2	Contig_5976	NW_011322503.1	3260	5771	+	104918638	psmg2	XP_010728750.1	262
tyrosine-protein phosphatase non-receptor type 2-like	Contig_5976	NW_011322503.1	8800	20329	-	104918639	LOC104918639	XP_010728753.1	392
E3 ubiquitin-protein ligase RNF19A-like	Contig_5976	NW_011322503.1	37022	47893	+	104918640	LOC104918640	XP_010728754.1	858
sperm-associated antigen 1A-like	Contig_5976	NW_011322503.1	52053	63784	-	104918661	LOC104918661	XP_010728782.1	392
DNA-directed RNA polymerases I, II, and III subunit RPABC4	Contig_5976	NW_011322503.1	66798	68311	-	104918641	polr2k	XP_010728756.1	58

Supplemental Table A5 continued

Protein Name	*RAD Locus*	*Scaffold*	*Start*	*Stop*	*Strand*	*Gene ID*	*Gene Name*	*Protein Accession*	*Length*
serine/threonine-protein kinase 3	Contig_5976	NW_011322503.1	68963	73628	+	104918642	stk3	XP_010728757.1	501
ribonuclease UK114-like isoform X1	Contig_5976	NW_011322503.1	76582	80750	+	104918643	LOC104918643	XP_010728758.1	135
60S ribosomal protein L30	Contig_5976	NW_011322503.1	96416	100475	+	104918646	rpl30	XP_010728760.1	116
lysosomal-associated transmembrane protein 4B-like	Contig_5976	NW_011322503.1	106763	116729	-	104918647	LOC104918647	XP_010728761.1	225
tumor protein D52 isoform X1	Contig_22948	NW_011322601.1	190120	207173	+	104922215	tpd52	XP_010733284.1	201
zinc finger protein 704 isoform X2	Contig_22948	NW_011322601.1	221698	265950	-	104922217	znf704	XP_010733287.1	521
phosphoprotein associated with glycosphingolipid-enriched microdomains 1	Contig_22948	NW_011322601.1	270484	286243	-	104922218	pag1	XP_010733289.1	452
39S ribosomal protein L53, mitochondrial	Contig_22948	NW_011322601.1	327674	329817	+	104922220	mrpl53	XP_010733290.1	106
myelin P2 protein	Contig_22948	NW_011322601.1	331344	332570	+	104922221	pmp2	XP_010733291.1	134
protein FAM8A1	Contig_22948	NW_011322601.1	346760	350872	+	104922222	fam8a1	XP_010733292.1	315
protein phosphatase 1 regulatory subunit 36	Contig_10413	NW_011323196.1	31302	38217	-	104931679	ppp1r36	XP_010745046.1	456
thiopurine S-methyltransferase isoform X1	Contig_10413	NW_011323196.1	38925	42214	-	104931680	tpmt	XP_010745048.1	263
exostosin-1c	Contig_10413	NW_011323196.1	55497	134499	+	104931682	LOC104931682	XP_010745051.1	623
isocitrate dehydrogenase	Contig_10413	NW_011323196.1	134502	141801	+	104931681	idh1	XP_010745050.1	414
homeobox protein PKNOX2	Contig_14312	NW_011323104.1	329703	359373	-	104930637	pknox2	XP_010743787.1	483
transmembrane protein 218	Contig_14312	NW_011323104.1	405468	417594	+	104930639	tmem218	XP_010743791.1	90
roundabout homolog 4	Contig_14312	NW_011323104.1	425862	444747	+	104930640	robo4	XP_010743792.1	1130
roundabout homolog 3-like isoform X2	Contig_14312	NW_011323104.1	455457	462905	-	104930638	LOC104930638	XP_010743790.1	266
roundabout homolog 2-like	Contig_14312	NW_011323104.1	463140	495644	-	104930641	LOC104930641	XP_010743793.1	1025
roundabout homolog 2-like	Contig_14312	NW_011323104.1	495788	532651	-	104930642	LOC104930642	XP_010743794.1	124
alpha-aminoadipic semialdehyde synthase, mitochondrial-like, partial	Contig_7751	NW_011323997.1	211	9890	+	104937320	LOC104937320	XP_010751829.1	481
receptor-type tyrosine-protein phosphatase zeta	Contig_7751	NW_011323997.1	14338	57972	-	104937321	ptprz1	XP_010751830.1	1373
alpha-aminoadipic semialdehyde synthase, mitochondrial, partial	Contig_31777	NW_011322893.1	283	7862	-	104927983	aass	XP_010740489.1	459
fez family zinc finger protein 1	Contig_31777	NW_011322893.1	13585	19365	-	104927984	fezf1	XP_010740490.1	437
calcium-dependent secretion activator 2	Contig_31777	NW_011322893.1	28785	260684	-	104927985	cadps2	XP_010740491.1	1295
PC2									
neurexin-2-like	Contig_22921	NW_011322530.1	163264	483252	+	104919663	LOC104919663	XP_010730030.1	1061
neurobeachin-like	Contig_37123	NW_011322641.1	414054	606100	-	104923311	LOC104923311	XP_010734670.1	2884
protein mab-21-like 1	Contig_37123	NW_011322641.1	467423	468502	+	104923291	mab21l1	XP_010734646.1	359
stAR-related lipid transfer protein 13-like isoform X1	Contig_37123	NW_011322641.1	610376	657531	-	104923292	LOC104923292	XP_010734647.1	1153
melanocortin receptor 4-like	Contig_14586	NW_011323077.1	130952	131938	+	104930272	LOC104930272	XP_010743320.1	328
protein FAM19A2-like	Contig_21372	NW_011322498.1	124714	159521	+	104918365	LOC104918365	XP_010728390.1	132
monocarboxylate transporter 1-like	Contig_21372	NW_011322498.1	279892	283931	+	104918366	LOC104918366	XP_010728391.1	456
synaptonemal complex protein 1 isoform X1	Contig_21372	NW_011322498.1	286759	292231	+	104918367	sycp1	XP_010728392.1	720
mucin-5B-like isoform X1	Contig_21372	NW_011322498.1	292846	294335	-	104918368	LOC104918368	XP_010728396.1	273
sodium-coupled monocarboxylate transporter 1-like	Contig_21372	NW_011322498.1	295455	302706	-	104918370	LOC104918370	XP_010728399.1	605
thyrotropin subunit beta	Contig_21372	NW_011322498.1	304788	305555	+	104918371	tshb	XP_010728401.1	146

Supplemental Table A5 continued

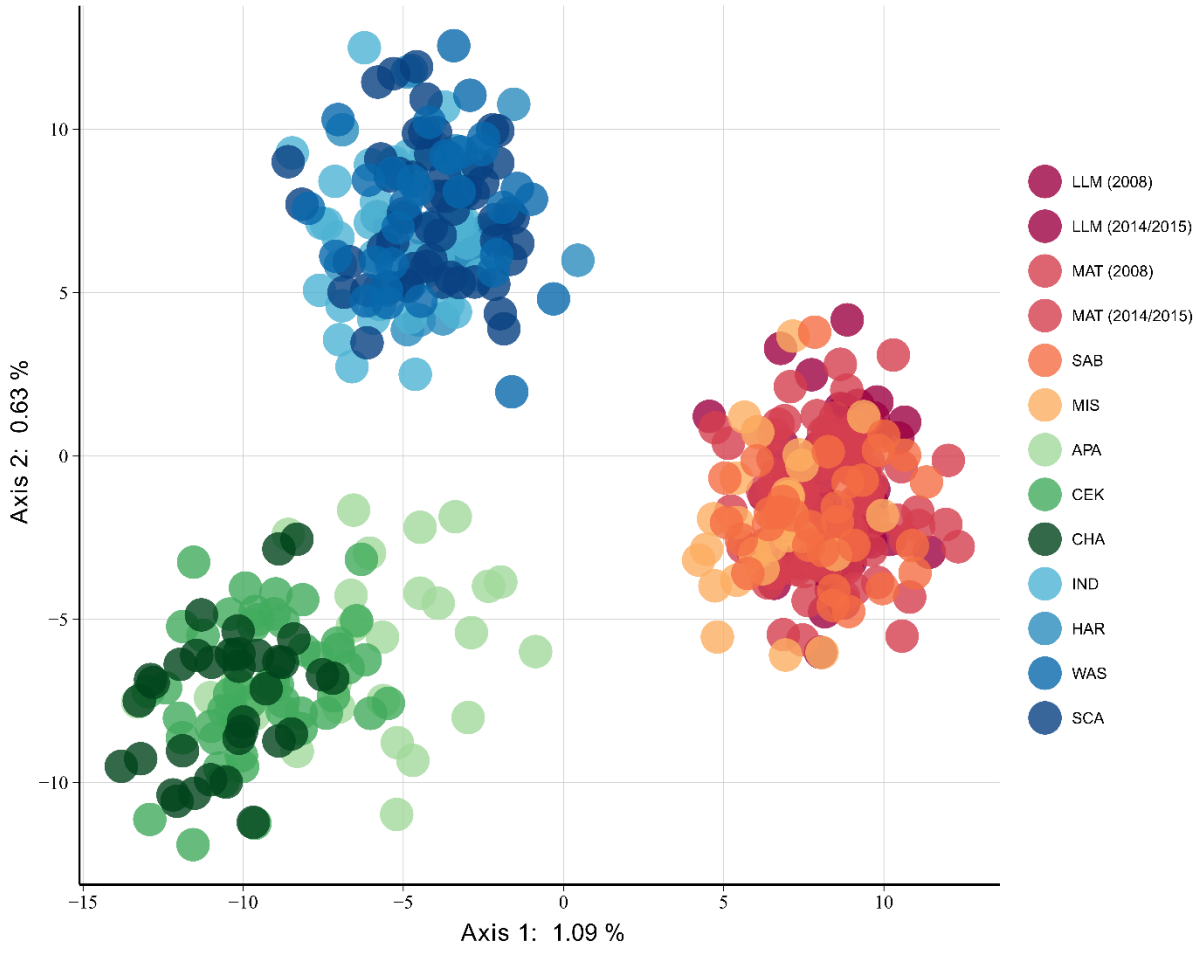
Protein Name	*RAD Locus*	*Scaffold*	*Start*	*Stop*	*Strand*	*Gene ID*	*Gene Name*	*Protein Accession*	*Length*
UPF0489 protein C5orf22 homolog	Contig_64456	NW_011323556.1	2230	13248	+	104934799	LOC104934799	XP_010748847.1	447
uncharacterized protein LOC104934800	Contig_64456	NW_011323556.1	18997	24937	+	104934800	LOC104934800	XP_010748848.1	306
actin-related protein 2/3 complex subunit 5-like	Contig_64456	NW_011323556.1	31981	36067	+	104934801	LOC104934801	XP_010748849.1	150
E3 ubiquitin-protein ligase RING2	Contig_64456	NW_011323556.1	42062	46463	-	104934803	rnf2	XP_010748850.1	342
ADAMTS-like protein 2	Contig_64456	NW_011323556.1	49815	67932	+	104934805	LOC104934805	XP_010748852.1	1021
glycogen debranching enzyme	Contig_64456	NW_011323556.1	80695	108664	-	104934804	agl	XP_010748851.1	1542
neurotrimin-like	Contig_12224	NW_011323104.1	19286	160062	-	104930635	LOC104930635	XP_010743785.1	343
myotubularin-related protein 4 isoform X1	Contig_19050	NW_011322948.1	67100	101737	-	104928772	mtmr4	XP_010741419.1	1224
4-hydroxyphenylpyruvate dioxygenase	Contig_19050	NW_011322948.1	110909	117837	+	104928774	LOC104928774	XP_010741422.1	395
heat shock factor protein 5	Contig_19050	NW_011322948.1	207257	209076	-	104928775	hsf5	XP_010741423.1	455
transcription elongation factor SPT4	Contig_19050	NW_011322948.1	216520	219462	+	104928780	supt4h1	XP_010741431.1	86
peripheral-type benzodiazepine receptor-associated protein 1	Contig_19050	NW_011322948.1	234481	291269	-	104928781	bzrap1	XP_010741433.1	2208
transmembrane protein 81	Contig_13928	NW_011323399.1	224792	226038	-	104933568	tmem81	XP_010747363.1	284
ras-related protein Rab-7L1	Contig_13928	NW_011323399.1	230044	234337	+	104933566	rab29	XP_010747361.1	217
pseudopodium-enriched atypical kinase 1	Contig_33160	NW_011322606.1	54124	78310	-	104922377	peak1	XP_010733483.1	1942
sorting nexin-33	Contig_33160	NW_011322606.1	117444	136610	-	104922378	snx33	XP_010733484.1	561
snurportin-1 isoform X1	Contig_33160	NW_011322606.1	137841	145434	+	104922379	snupn	XP_010733485.1	389
tyrosine-protein phosphatase non-receptor type 9	Contig_33160	NW_011322606.1	153207	171853	+	104922380	ptpn9	XP_010733487.1	570
paired amphipathic helix protein Sin3a	Contig_33160	NW_011322606.1	182112	197482	+	104922381	sin3a	XP_010733488.1	1273
cellular retinoic acid-binding protein 1	Contig_33160	NW_011322606.1	215994	233119	-	104922382	crabp1	XP_010733490.1	137
WD repeat-containing protein 61	Contig_33160	NW_011322606.1	239002	244325	+	104922383	wdr61	XP_010733492.1	305
solute carrier family 25 member 44-like	Contig_33160	NW_011322606.1	246713	248766	-	104922385	LOC104922385	XP_010733494.1	317
iron-responsive element-binding protein 2	Contig_33160	NW_011322606.1	253484	267283	+	104922384	ireb2	XP_010733493.1	972
sorting nexin-1-like	Contig_33160	NW_011322606.1	274210	285368	-	104922386	LOC104922386	XP_010733496.1	807

Supplemental Table A6 List of species with similar patterns of genetic divergence. Gulf/Atlantic refers to species for which significant genetic differences have been observed between the Gulf of Mexico and the western Atlantic Ocean; Western Gulf/Eastern Gulf refers to species for which significant genetic differences have been observed between the western and eastern Gulf of Mexico.

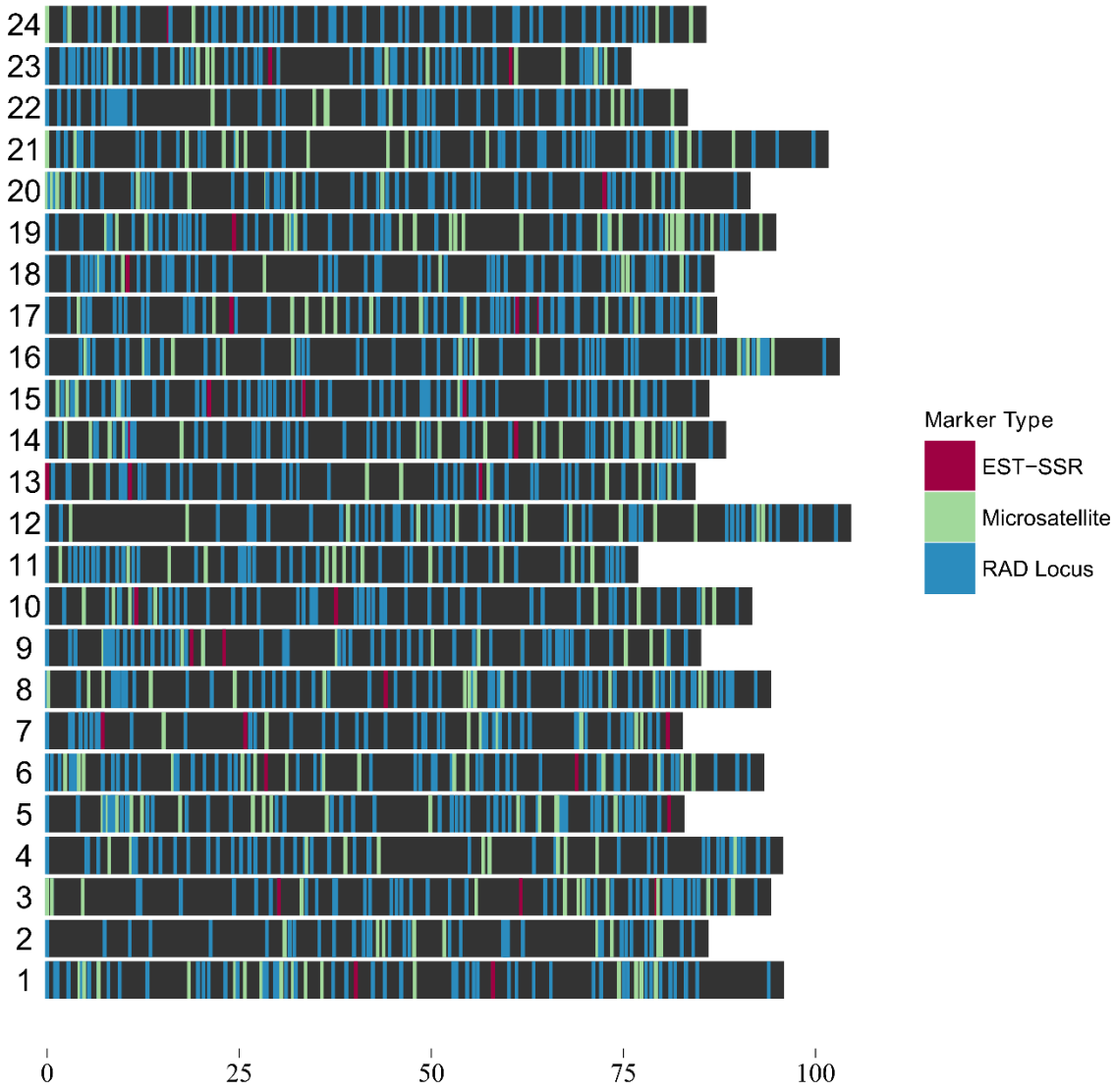
Common name	Scientific Name	Family	Life History/Habitat Preference	Citation
Gulf/Atlantic				
black drum	*Pogonias cromis*	Sciaenidae	Coastal demersal	Leidig *et al.*, 2015
black sea bass	*Centropristis striata*	Serranidae	Reef associated	Bowen and Avise, 1990
menhaden	*Breviorta tyrannus / B. patronus*	Clupeidae	Coastal pelagic	Bowen and Avise, 1990
Atlantic sturgeon	*Acipenser oxyrhynchus*	Acipenseridae	Anadromous demersal	Bowen and Avise, 1990
greater amberjack	*Seriola dumerili*	Carangidae	Coastal Pelagic/Reef associated	Gold and Richardson, 1998
king mackerel	*Scomberomorus cavalla*	Scombridae	Coastal pelagic	Gold *et al.*, 2002
gray snapper	*Lutjanus griseus*	Lutjanidae	Reef associated	Gold *et al.*, 2009
southern flounder	*Paralichthys lethostigma*	Paralichthyidae	Coastal demersal	Anderson *et al.*, 2012
blacknose sharks	*Carcharhinus acronotus*	Carcharhinidae	Coastal demersal	Portnoy *et al.*, 2014
bonnethead sharks	*Sphyrna tiburo*	Sphyrnidae	Coastal demersal	Portnoy *et al.*, 2015
Western Gulf/Eastern Gulf				
lane snapper	*Lutjanus synagris*	Lutjanidae	Reef associated	Karlsson *et al.*, 2009
blacknose sharks	*Carcharhinus acronotus*	Carcharhinidae	Coastal demersal	Portnoy *et al.*, 2014

APPENDIX B

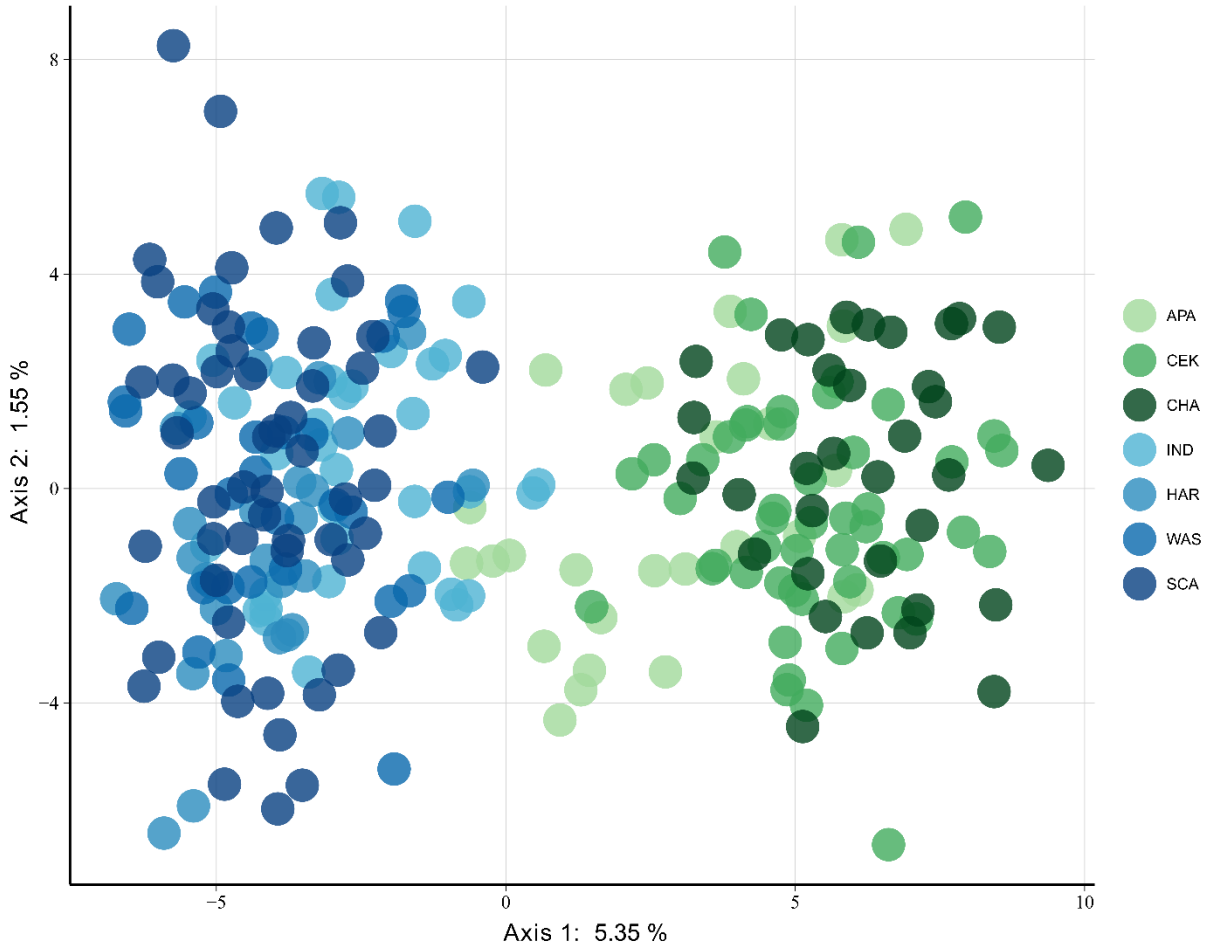
SUPPLEMENTAL FIGURES



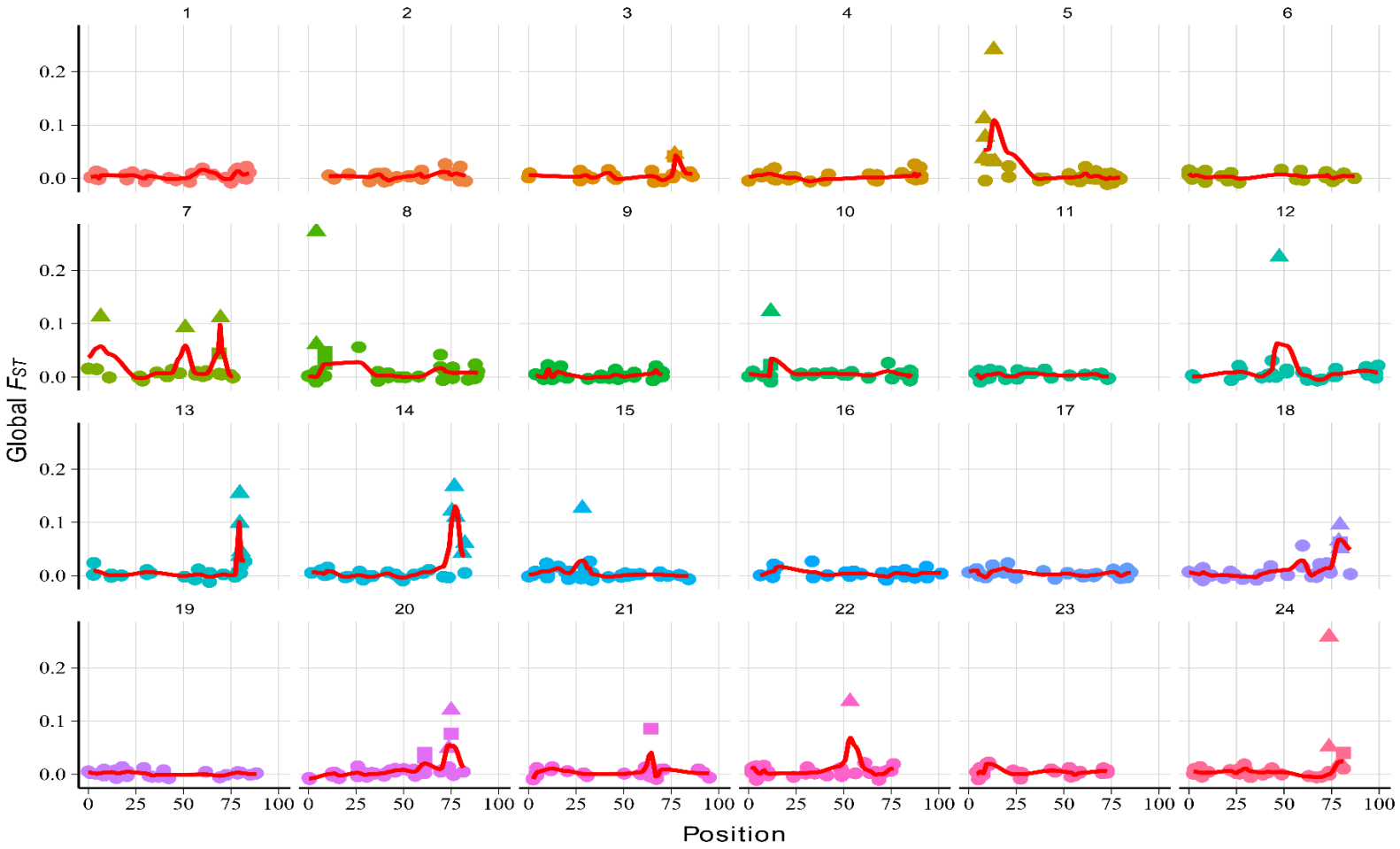
Supplemental Figure B1: Principal components analysis (PCA), using all loci (*n* = 1,539), and including two pairs of temporal samples from Lower Laguna Madre (LLM) and West Matagorda Bay (MAT).



Supplemental Figure B2 Consensus linkage map for red drum, including data from two mapping crosses. Each grey, horizontal bar represents one of 24 red drum linkage groups (y-axis), and colored vertical bars represent genetic markers at a particular position (x-axis) on the linkage group. Markers are colored by type: blue = haplotyped RAD loci (containing one or more SNP/indel loci); green = anonymous microsatellite loci; red = EST-derived (Type-I) microsatellite loci (EST-SSRs).



Supplemental Figure B3 Principal components analysis (PCA), using outlier loci (*n* = 146), and including samples from eastern Gulf of Mexico (Apalachicola: APA, Cedar Key: CEK, Charlotte Harbor: CHA) and Atlantic Ocean (Indian River: IND, Hampton River: HAR, Wassaw Sound: WAS, South Carolina: SCA) regions.



Supplemental Figure B4 A plot of global F_{ST} for each locus, organized by position on the 24 red drum linkage groups. Each panel represents a single linkage group, with global F_{ST} displayed on the y-axis and position on the linkage group (centiMorgans) displayed on the x-axis. The red line on each panel is a smoothed mean of F_{ST} for the group, calculated with a loess function, as implemented in the ggplot2 package in R.

APPENDIX C

USING LINKNE TO CALCULATE EFFECTIVE POPULATION SIZE

The program LINKNE extends the method of calculating contemporary effective population size (N_e) from linkage disequilibrium (LD) by using unphased genotype data (Waples and Do, 2008) to incorporate the effects of linkage and (i) estimate N_e in past generations, and (ii) remove bias caused by violating the assumption that all pairs of loci in the dataset are unlinked.

Following Waples (2006), the observed LD, as measured by r^2, for a given pair of loci can be broken down into two components: one due to the effects of genetic drift in finite populations, and one due to the effects of sampling a limited number of individuals from the population. Thus:

$$E(\hat{r}^2_{\text{total}}) = E(\hat{r}^2_{\text{drift}}) + E(\hat{r}^2_{\text{sample}}) \quad (1)$$

Effective population size (N_e) can be estimated by considering the effects of the component of LD due to drift alone, which can be obtained by rearrangement:

$$E(\hat{r}^2_{\text{drift}}) = E(\hat{r}^2_{\text{total}}) - E(\hat{r}^2_{\text{sample}}) \quad (2)$$

Both values on the right side of the above equation can be estimated from observed genotype data.

Estimating LD Component Due to Sampling Variation

Weir and Hill (1980) showed that for a randomly mating population, the contribution to LD of sampling a finite number of individuals could be estimated as

$$E(\hat{r}^2_{\text{sample}}) = \frac{1}{S} \tag{3}$$

where S is the number of individuals sampled. However, England *et al.,* (2006) found a large downward bias in estimates of effective size when using this equation, particularly if sample size was small relative to the true N_e. To address this bias, Waples (2006) suggested a bias correction, based on simulated data, which depended on the sample size: For S > 30,

$$E(\hat{r}^2_{\text{sample}}) = \frac{1}{S} + \frac{3.19}{S^2} \tag{4}$$

and for S < 30:

$$E(\hat{r}^2_{\text{sample}}) = 0.0018 + \frac{0.907}{S} + \frac{4.44}{S^2} \tag{5}$$

Following Waples and Do (2008), r^2_{sample} is averaged across pairs of alleles and loci. To account for effects of different sample sizes (due to missing data) and number of alleles, a weighting factor is applied to each locus pair. The weight of locus pair (*i*, *j*) is calculated as

$$w_{ij} = n_{ij} S_{ij}^2 \tag{6}$$

where…

$$n_{ij} = (n_i - 1)(n_j - 1)$$

and n_i and n_j are number of alleles at loci i and j, respectively, and S_{ij} is the sample size of individuals genotypes at each locus.

The weighted arithmetic mean across loci is then calculated as…

$$\hat{r}^2_{\text{sample}} = \frac{1}{W} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} r^2_{ij,sample} \tag{7}$$

where…

$$W = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij}$$

Estimating Total LD

For each pair of alleles in pairwise comparisons of loci in the dataset, $\hat{r}^2_{\text{total}}$ is estimated using Burrow's composite linkage disequilibrium measure (*D*), as in Weir (1996), with the formula:

$$D = \hat{\Delta}_{AB} = \frac{n_{AB}}{n} - 2\hat{p}_A\hat{q}_B \tag{8}$$

where…

$n =$ number of individuals genotyped at both loci,

$n_{AB} = 2n_1 + n_2 + n_3 + \frac{n_4}{2}$,

$n_1 =$ count of double homozygous individuals,

n_2 and $n_3 =$ counts of individuals heterozygous for one or the other allele,

$n_4 =$ count of double heterozygous individuals, and

$\hat{p}_A$ and $\hat{q}_B =$ allele frequencies of each allele among individuals genotyped at both loci.

Burrow's *D* can be standardized by allele frequency to produce the correlation coefficient $\hat{r}_{AB}$…

$$\hat{r}_{AB} = \frac{\hat{\Delta}_{AB}}{\sqrt{(\hat{p}_A(1-\hat{p}_A) + (\hat{h}_{AA} - \hat{p}_A^2)(\hat{q}_B(1-\hat{q}_B) + (\hat{h}_{BB} - \hat{q}_B^2))}} \quad (9)$$

where…

$\hat{h}_{AA}$ and $\hat{h}_{BB}$ = the observed proportions of homozygous genotypes of each allele in the individuals genotyped at both loci, and

$\hat{p}_A$ and $\hat{q}_B =$ allele frequencies of each allele among individuals genotyped at both loci.

The square of this value ($\hat{r}^2_{AB}$) is calculated for each locus pair (averaging across all allelic combinations), weighted as in the previous section, and averaged across all pairs of loci as…

$$\hat{r}^2_{\text{total}} = \frac{1}{W} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} r^2_{ij,total} \tag{10}$$

Estimating N_e with LD from Markers with Known Linkage Relationships

Binning Estimates

To obtain precise estimates of N_e at multiple points in time, estimates from multiple locus pairs must be binned together across a range of recombination rates. Hayes *et al.*, (2003) derived an approximate relationship between recombination rate and time as…

$$t = \frac{1}{2c} \tag{11}$$

where…

t = the number of generations in the past to which the estimate applies, and

c = the recombination rate between a pair of loci (in Morgans)

Assuming this relationship holds for all possible recombination rates, this means that the majority of the spectrum of recombination rates reflects N_e at very recent generations. For example, while $t = 1$ when loci are completely unlinked ($c = 0.5$), t does not reach two generations in the past until $c = 0.25$ and does not reach ten generations in the past until $c = 0.1$. Because of this, binning locus pairs into equally sized windows, based on recombination rates, produces trend lines that are not very informative as they mostly reflect recent N_e and tend to collapse past estimates into a small region at the end of the trend line. A more informative way of binning locus pairs is by generation, which allows finer scale changes to be revealed but requires that bins are larger at larger recombination rates. LINKNE allows users to choose whether bins should be defined by equally sized recombination rate windows or by generations. If generation-based bins are specified, the program first produces recombination rate-based bins (based on a size specified by the user) and then iteratively merges bins when the bin midpoints refer to time points within two generations of each other (based on the above equation).

Calculating N_e for each bin

As stated above, given $\hat{r}^2_{\text{total}}$ and $\hat{r}^2_{\text{sample}}$, the component of LD due to drift can be calculated by Equation 2. Weir and Hill (1980) showed that under the assumption of random mating and if N_e and S are relatively large, r^2-drift can be written as a function of N_e and the recombination rate (c, in Morgans) between pairs of loci…

$$\hat{r}^2_{\text{drift}} = \frac{(1-c)^2 + c^2}{2N_e c(2-c))} \tag{12}$$

Hill (1981) simplified this equation by separating the term relating to recombination rate from the term relating to N_e…

$$\hat{r}^2_{\text{drift}} = \frac{\gamma}{N_e} \tag{13}$$

where…

$\gamma = \frac{(1-c)^2+c^2}{2c(2-c)}$, and

c = the mean recombination rate of a bin.

This is rearranged to calculate N_e as…

$$N_e = \frac{\gamma}{\hat{r}^2_{\text{drift}}} \tag{14}$$