Genetic Resources for Improved Selection and Management of Aquaculture and Conservation Fish Species

by

Honggang Zhao

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 14, 2019

Keywords: Florida bass, white bass, walleye,
SNP, GBS, MassARRAY

Approved by

Eric Peatman, Chair, Professor of Fisheries, Aquaculture, and Aquatic Sciences
Benjamin Beck, Research director of USDA-ARS Aquatic Animal Health Research Unit
Scott McElroy, Professor of Crop Soil and Environmental Sciences
Tonia Schwartz, Assistant Professor of Biological Science

Abstract

Molecular markers are versatile tools for conservation genetics allowing the identification of populations in genetic crisis, resolving taxonomic uncertainties, and establishing management units within fish species. Advances in biotechnologies have enabled SNP discovery and genotyping in a cost-effective and parallel manner. At the same time, huge progress has been achieved for high-throughput SNP genotyping and marker applications thanks to the introduction of assay-based genotyping platforms. Here, I describe SNP discovery and panel development in three key aquatic species in the southeastern United States: Florida bass (*Micropterus floridanus*), white bass (*Morone chrysops*), and walleye (*Sander vitreus*).

Florida bass are arguably the most important freshwater sportfish in North America. In my first study, I carried out genotyping-by-sequencing (GBS) and developed SNP panels for Florida bass parentage assignment. Parentage analysis demonstrated that the developed marker panels were capable of accurate parentage assignment and are more powerful than existing microsatellite tools for the species. The SNP resources created in this study should facilitate parentage-based research and breeding, genetic tagging, and conservation of Florida bass.

White bass are temperate bass species with high commercial and ecological value in North America. Population structure analyses using GBS data revealed two distinct genetic clusters among a domesticated white bass line and five potential founder stocks. Additionally, a 57-SNP assay was successfully developed to assign bass individuals back to their origin populations. The developed panels should augment ongoing efforts toward white bass conservation and selective breeding.

Walleye are ecologically and economically important fish found in freshwater river and lake systems in North America. Southern walleye in Mobile River Basin were previously identified as a long-isolated lineage that genetically diverged from other walleye populations in North America. Here, I utilized GBS data to infer genetic diversity and structure among northern and southern walleye populations. Additionally, a SNP assay with 68 diagnostic markers was developed for rapid and accurate identification of genetic purity and classification of various hybrid classes. The availability of high-quality GBS datasets and a large set of diagnostics SNPs should greatly facilitate conservation and population genomics studies in this key species.

Acknowledgments

I could not finish this dissertation work without the help and guidance of my supervisor, Dr. Eric Peatman, and the innumerable opportunities he provided me in reaching my goal of pursuing research in fish genetics and genomic sciences. I would also like to thank all my committee members, Dr. Benjamin Beck, Dr. Scott McElroy, Dr. Tonia Schwartz, Dr. Cova Arias, and my university reader, Dr. Charles Chen. They gave me invaluable advice and assistance throughout the entirety of my dissertation work.

I must thank my fellow lab members whom I have shared many memories with, Dr. Chao Li, Dr. Wilawan Thongda, Dr. Dongdong Zhang, Dr. Haitham Mohammed, Spencer Gowan, Taylor Brown, Yupeng Luo, and Lauren Davis. I am also thankful to my current lab mates: Dr. Katherine Silliman, Matt Lewis, Sarah Johnson, Aaron Fewell, and Megan Justice. Helping each other with the lab work across many projects helped us build strong relationships as colleagues and friends. Furthermore, I would like to thank Milla Kaltenboeck and Dr. Huseyin Kucuktas.

I am thankful to the China Scholarship Council for granting me the scholarship throughout my Ph.D. program at Auburn University. I am also grateful to Dr. Peatman for providing funding during my additional half year so that I could complete the dissertation projects.

Finally, I would like to thank all of my friends and family for their love and support. They have always stayed beside me and believed in me. Their unconditional love, deep understanding, and encouragement are the most precious things in my life.

Table of Contents

List of Tables

List of Figures

Chapter I Introduction and Literature Review

1. Advances in Aquaculture and Conservation Genetics

Driven by the biotechnology revolution and human biomedical development, genomic sciences have made drastic advances in the last ten years [1]. It is not just the high-throughput sequencing that has revolutionized the way scientific work is conducted; the reduced cost of sequencing and the availability of various bioinformatic tools have made these technologies applicable to all aspects of molecular research across different organisms, including fish species important to aquaculture and conservation. Aquaculture is the fastest-growing sector of agriculture and is expected to maintain its rapid growth in the coming decades in the facing of population increase and declining wild-catch fisheries [2]. The advance in genomic techniques will be most valuable in addressing questions related to aquaculture, as the use of genomics and selective breeding in aquaculture generally lags behind plant and farm animal industries [3]. Towards this end, various genome-based approaches have been used for stock enhancement in aquaculture species, including traditional methods such as polyploidization, gynogenesis, androgenesis or sex reversal, as well as the new technologies of marker-assisted selection (MAS), genome selection (GS), and genome editing [4]. On the other hand, the multi-faceted genomic approaches are currently being used to address conservation issues, including but not limited to estimating neutral population parameters (e.g. effective population size, population diversity, and population differentiation), understanding the genetic basis of local adaptation or inbreeding depression, and further predicting a population's viability to adapt to climate change based on genomic information [5]. Nevertheless, the application of genomics in aquaculture and conservation genetic studies has been limited to a relatively few species deemed important due to their impact on the economy and/or and a threatened/imperiled status. This is due, in large part, to a paucity of modern, verified genetic markers in non-model fish species.

2. Modern Genetic Marker: Single Nucleotide Polymorphism

All organisms are subject to mutation or alteration in genome loci as a result of interactions with the environment, leading to heritable genetic variation discernable within and among individuals, species, and higher-order taxonomic groups [6]. Genetic marker technologies can be used to reveal these variations. Several marker types have been used in aquaculture and conservation genetic studies, including allozyme, mitochondrial DNA (mtDNA), restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLPs), microsatellites, expressed sequence tags (ESTs) and single nucleotide polymorphisms (SNPs) [6]. Among them, allozymes are one of the earliest markers used in aquaculture genetics, with their polymorphism associated with allelic variants within proteins (enzymes) [7-9]. Although allozymes have been widely utilized for early conservation (e.g. [10-13]) and aquaculture (e.g. [14-16]) research, the low level of genetic resolution, required sacrifice of the organisms, and issues related to silent or synonymous substitutions (i.e. allozymes with identical electrophoretic mobility may mask underlying sequence variation) have limited the usage of allozymes [6]. RFLPs are markers that detect genetic variation based on different DNA fragment lengths generated by restriction endonucleases [17]. This marker type has been broadly employed in aquaculture and conservation studies in the 1980s (e.g. [18-20]) but constraints such as low levels of polymorphism and the required sequence information for target loci have limited their usage in fish species lacking known molecular information [6]. RAPD is a genetic marker that using 8-10 arbitrary primers to randomly amplify anonymous segments of nuclear DNA [21, 22]. The advantages of this marker type include cost-effectiveness (with primers commercially available) and no requirement for known targeted sequences [6]. However, RAPDs are inherited as Mendelian markers in a dominant fashion, making it difficult to distinguish between homozygotes and heterozygotes [6]. In addition, their low reproducibility due to the low annealing temperature used in PCR has limited the use of this marker [23]. Later-generation marker types such as AFLP and microsatellite are generally neutral markers and can provide better informative content and a higher resolution of genetic variation compared with RFLP and RAPD [6]. Similar to RFLP, AFLP is a technique that targets genetic variation associated with restriction sites [24]. The unique feature of this method is the addition of adaptors of known sequences for PCR amplification, which makes it a PCR-based multi-locus fingerprinting technique [24]. The major strengths of the AFLP include large number of revealed polymorphisms, high reproducibility, and moderate costs. However, the need for special equipment for electrophoretic analysis and the poor genetic information on a per marker basis hampered the spread of the AFLP method to studies of fish species [25]. Microsatellites are tandem repeats of 2-6 bp simple sequences caused by polymerase slippage during DNA replication [26, 27]. This marker type enjoys relatively high abundance in genomes, is evenly distributed, and has a high rate of polymorphism [6]. Because of these properties, microsatellites have been extensively applied to genetic investigations in aquatic species, including genome mapping, parentage or kinship determination, population structure identification, and hybridization analyses [28]. Although microsatellites have been a popular genetic marker over the past two decades, the low multiplexing capacity, low throughput, and the required obsolescence of ABI sequencer for scoring have contributed to a transition to SNP markers over the last few years.

3. Molecular Basis of SNP

Single nucleotide polymorphisms (SNPs) are a marker type generated by point mutations that give rise to alternative alleles at a given nucleotide position within a locus [6]. Such point mutations have been well-described since the early days of DNA sequencing, but it was only in the late 1990s that scientists could genotype SNPs using gene chip technology [6]. Theoretically, each SNP within a locus could produce up to four alleles (A, T, C, G). Practically, however, SNPs are generally bi-allelic markers that are restricted to one of the two alleles (e.g. C/T, A/G). SNPs are ubiquitously located in the genomes in both coding (exons) and non-coding (e.g. introns, 5’UTR, and 3’UTR) regions, where SNPs in the coding regions can be divided into synonymous and nonsynonymous SNPs. Compared to multiple-allele markers like microsatellites, SNPs are less polymorphic on per marker basis, but this shortcoming can be balanced by their great abundance [6]. SNPs have rapidly gained the center stage of aquaculture and conservation genetic research due to their abundance in the genome (both coding and non-coding regions), low cost of genotyping, low genotyping error rate, ease of multiplexing, great reproducibility, amenability to high throughput processes, and their high level of resolution [29-32]. SNPs have been broadly applied to aquaculture and aquatic conservation studies in recent years for various purposes such as the determination of the population structure, population genomics, population or species identification, hybridization rates, and migratory dynamics [42-49].

4. SNP Discovery and Genotyping

Two principle steps are needed for the use of SNPs: marker discovery and genotyping [30]. SNP discovery is the process of identifying polymorphic sites in the genome of a species and/or populations of interest [30]. Before the advent of high-throughput sequencing, SNP discovery was often expensive, labor-intensive, and could potentially introduce ascertainment bias through the use of limited sample numbers. Initially, Sanger sequencing was the workhorse for *de novo* genome sequencing and the discovery of SNP markers [30]. Although Sanger sequencing has been used successfully to generate SNPs for many non-model aquatic species [33-35], it is not practical to use this technique for SNP discovery across multiple loci and a large number of individuals, as the process would be prohibitively expensive and time-consuming. Another former method of SNP discovery is through expressed sequence tags (ESTs), which are sequences of cDNA from processed mRNA [36]. In 2007, Hayes et al. (2007) identified 2,507 putative SNPs from the alignment of Atlantic salmon ESTs and successfully built a genetic linkage map based on these EST-derived SNPs [37, 38]. Following their pioneering work in Atlantic salmon, similar methods have been applied to SNP discovery in many other aquatic species, including catfish [39], common carp [40], grass carp [41], olive flounder [42], turbot [43], and Pacific oyster [44]. However, the use of ESTs for SNP discovery may confound neutral parameter estimates in population genetic studies, as the EST-based discovery process only includes variation in coding regions [29]. One of the principal challenges that are of most significance in the SNP discovery process is ascertainment bias, as failure to minimize ascertainment bias may lead to the introduction of uninformative markers. Unfortunately, for typical Sanger-based or EST-based SNP discovery, only few individuals are selected for the discovery process when the cost per sample and labor are taken into account [29].

4.1 SNP Discovery

The discovery of SNPs has inextricably linked to advances in next-generation sequencing (NGS) and the development of bioinformatics tools over the last decade. For aquatic species lacking reference genome resources, several approaches combining reducing genome complexity and cost-effective NGS have been proposed for SNP discovery, including the transcriptome sequencing of a pool of individuals (RNA-seq, e.g. [45, 46]) and several restriction site-associated DNA sequencing approaches, such as RAD-seq [47], genotyping-by-sequencing or GBS [48], 2b-RAD [49], double-digest RAD-seq or ddRAD-seq [50], and ezRAD [51]).

4.1.1. RNA-seq

RNA-seq can reduce genome complexity by sequencing only the transcriptomes of a pool of individuals [52-54]. Using RNA-seq data, SNP discovery and gene expression analyses can be conducted simultaneously. The basic procedures for converting raw transcriptome sequences into SNP results including initial raw reads processing and trimming, *de novo* assembly or preparation of a reference genome, read mapping, and final SNP calling [55]. Various software, either commercial or open-source, are available for these procedures. Assemblers such as CLC Genomics Workbench (CLC Bio, Aarhus, Denmark),), Newbler (gsAssembler, 454 Life Sciences, Roche Diagnostics), Alpheus [56], Trinity [57], Mira [58], AbySS [59], Velvet [60], and Oases [61] can be used for *de novo* transcriptome assembly. For species with a reference genome, short reads from RNA-seq data could be mapped directly onto the reference sequence through existing bioinformatic tools such as CLC Workbench, SSAHA [62], BLAT [63], MAQ [64], SeqMap [65], SOAP [66], Bowtie [67], BWA [68], Samtools [69], and Bedtools [70]. Subsequently, SNP callers such as CLC Workbench, Genome Analysis Toolkit (GATK) [71], Samtools/BCFtools [69], freebayes [72], and SOAPsnp [66] are available for variant calling. RNA-seq can produce large numbers of SNPs at relatively low costs, and therefore has been applied in a multitude of aquaculture and conservation studies. For example, Baranski et al. (2014) identified a total of

473,620 putative SNPs/indels from RNA-seq data in India black tiger shrimp (*Penaeus monodon*) populations and developed an Illumina iSelect genotyping array containing 6,000 SNPs for the genotyping of additional tiger shrimp samples [73]. With these SNP resources, they built a high-density genetic map for this specific aquaculture species and highlighted its possible roles in future functional studies (e.g., QTL mapping, marker-assisted selection). Another RNA-seq study using pooled rainbow trout samples from a population selected for improved growth versus slow-growing cohorts identified 361 SNPs putatively associated with growth traits [74]. In this study, although the libraries from pooled samples were sequenced at low depth of coverage (~0.97X), approximately 70% success rate in detecting polymorphic/true SNPs was achieved through the MassARRAY System (Agena Bioscience), suggesting the higher accuracy rate of RNA-Seq in SNP discovery [74]. These example studies clearly demonstrated that RNA-seq enables the simple identification of thousands of SNPs in a cost-effective and accurate manner.

4.1.2. Restriction Site-Associated DNA Sequencing

A range of techniques that rely on restriction site-associated DNA sequencing have emerged in the last few years, including genotyping-by-sequence or GBS [48], RAD-seq [47], 2b-RAD [49], ddRAD-seq [50], and ezRAD [51]. Unlike whole genome sequencing, the core feature of these techniques is to utilize restriction enzymes (REs) to cut the entire genome into DNA fragments and generate information on population-level variation. With the help of high-throughput sequencing and sample-specific barcodes, the sequencing data adjacent to the restriction cut sites can be easily obtained and used to call hundreds or thousands of genome-wide SNPs [75]. With these approaches, tens or hundreds of individuals that represent the breadth of the geographic distribution can also be included in a discovery panel on the NGS instruments, even in species with little or no previous genomic information s [29, 76-79].

The restriction-site-associated DNA sequencing techniques differ in the enzymes used for digestion, adaptor ligation utilized for library construction, the barcodes used for sample identification, and different size selection strategies applied for sequencing [80]. For instance, 2b-RAD employs the type IIB restriction enzymes for genomic DNA digestion. This unique type of enzyme can cut the DNA at their recognition sites and generate short fragments that are of equal size across (33–36 bp), thus omitting the size selection step [49]. Original RAD-seq is another technique that does not use size selection to reduce the set of loci to be sequenced; instead, a mechanical shearing step and the use of second adapter, or "Y" adapter, is required to produce fragments appropriate for Illumina sequencing [47, 80]. Among these techniques, GBS is a simple, cost-effective, and highly multiplexed approach that was initially developed for SNP identification and genotyping in crop genomes and populations [48]. In comparison with other techniques, the major advantages of GBS over other protocols are its technical simplicity (e.g., without direct size selection) and low cost. Additionally, a bioinformatics pipeline, TASSEL, is well-optimized for typical GBS data processing and analyzing [81]. Consequently, GBS has become one of the most widely used genomic approaches for high-throughput SNP discovery and genotyping in aquaculture, conservation, and evolutionary studies of non-model organisms [80].

The post-sequencing analyses share several basic steps for data processing and SNP calling, including raw data trimming, reads de-multiplexing (based on barcodes), and filtering reads based on the presence of the expected RE cut site and sequence quality [80]. If a reference genome is available, loci can be identified by mapping trimmed reads to the reference. Otherwise, the *de novo* loci will be produced by clustering similar trimmed reads together. After the locus generation, genotyping will be conducted using maximum likelihood [82] or Bayesian approaches [83]. Several bioinformatics pipelines are available for all parts of the data analyses, such as TASSEL, Stacks [84], PyRAD [85], as well as other publicly available scripts and pipelines. Among them, Stacks is a command-line based program that uses various flexible modules to conduct data analyses, from initial raw read processing to ultimate SNP calling and genotyping, plus an additional module (*populations*) that helps to compute some population statistics as well as exporting a variety of standard output formats. PyRAD, or ipyrad, can conduct quality filtering and identify *de novo* or reference-based SNPs, with the advantage that it can produce specific datasets for phylogenetic analysis. TASSEL is a pipeline designed for the efficient processing of GBS sequence data into SNP genotypes. The unique feature of TASSEL is that it can separate SNP discovery and genotyping into two phases, which reduces potential ascertainment biases and makes the pipeline highly suitable for use in association mapping analyses for complex traits.

4.2 SNP Validation and Genotyping

Although RNA-seq and restriction site-associated DNA sequencing techniques make it possible to identify and genotype thousands of SNPs directly from sequencing data, they still share some sources of sequencing and genotyping errors [86]. For example, PCR duplicates can occur at high frequencies in GBS or RAD-seq data (e.g., 20-60% reads) [87-89]. PCR duplicates are DNA fragment clones that may cause over-amplification of one allele more than other alleles at a given locus, thus lead to downstream genotyping errors (e.g., heterozygotes appear as homozygotes). Although issues like PCR duplicates can be addressed through bioinformatics approaches (e.g., remove PCR products with identical lengths, restricted in original RAD-seq) or alternative library prep protocols (e.g., ezRAD with Illumina PCR-free kit, [51]), marker screening and validation are still required for further uses of the developed marker resources. SNP genotyping platforms can verify and validate SNPs produced from sequencing data and genotype additional samples with the same sets of SNPs. With these platforms, researchers can screen SNPs on a larger set of individuals, populations, or strains, and provide genotype data for aquaculture and conservation research.

A variety of SNP genotyping platforms are available for aquaculture and conservation genetics, including small to medium scale platforms (e.g., TaqMan™, KASPar, Fluidigm, and Agena MassARRAY platforms), and high-density array genotyping methods (e.g., Affymetrix SNP arrays, Illumina genotyping arrays) [90]. Among them, Taqman™ and KASPar are 5'exonuclease methods that require the reaction between DNA sample and assay (including locus-specific primers and allele-specific fluorescent probes) for SNP genotype generation [91]. Therefore, the cost per SNP genotype is relatively high compared to high-density array approaches. For the Fluidigm system, DNA samples and assay are combined and loaded into reaction chambers on a nanofluidics chip. Following thermal cycling and fluorescence detection, a SNP genotype is generated either using a BioMark HD [92, 93] or EPI genotyping system [94, 95]. The Fluidigm system has high flexibility in the number of SNPs and the number of samples to be genotyped. The development of Affymetrix Axiom genotyping technology (Axiom® SNP array) has made it possible to genotype large numbers of SNPs (e.g., 190K, 250K, or 690K SNPs) through high-density SNP chips [96-98]. The primary task of SNP array development is to obtain SNPs from multiple sources including gene-associated SNPs, anonymous SNPs, and inter-specific SNPs. Such high-density arrays are valuable resources for genome-wide association studies (GWAS), fine QTL mapping, linkage map construction, and whole genome-based selection [98].

The Agena MassARRAY system is a platform commonly utilized in medium-scale SNP assay development and genotyping. The primary procedure of this system is its utilization of a single extension primer to generate allele-specific products with distinct masses. The mass of each allele-specific product is determined by MALDI-TOF mass spectrometry and converted into genotypes using MassARRAY Typer 4 software [99]. This system has several attractive features for researchers desiring an SNP genotyping assay with modest multiplexing. For instance, it offers minimal assay setup costs (due to unmodified oligonucleotide primers) and rapid turn-around time, with a typical run finished within a day from DNA extraction to data analysis. The MassARRAY system has been increasingly utilized for SNP genotyping and validation in aquaculture and conservation genetics studies, as well as in the studies reported in this dissertation [100-105].

5. Application of SNP markers in Aquaculture and Conservation Genomics

Aquaculture continues to grow significantly as a food source around the world. To meet the increased demand from the aquaculture global market, DNA marker technologies are playing essential roles in the improvement of breeding programs, disease management, and other aquaculture genetics research [6]. Meanwhile, the ability to examine thousands of SNPs with relative ease has made it possible to answer questions in the conservation of aquatic species, including estimation of neutral population parameters, identification of the genetic basis of local adaptation, and evaluation of a population’s viability or capacity to adapt to climate change [5]. In this section, I will summarize the recent application of SNP markers in aquaculture and conservation genomics studies.

5.1 Genetic Identification

Genetic identification of species, strains, or hybrid individuals using SNP markers is often required in aquaculture and conservation studies. In aquaculture, monitoring of the interspecific and/or intraspecific hybrid production is a common practice for the appropriate management of fish in the farms. For example, Karlsson et al. (2011) [106] developed a panel of 60 SNPs that are diagnostic in identifying individual Atlantic salmon as being farmed or wild, as well as their first-generation hybrids (F1). SNPs are also useful to determine the source origin of individuals escaped from aquaculture facilities. Using a higher-density Atlantic salmon SNP array, Pritchard et al. (2016) [107] identified a set of 200 SNPs that could differentiate an important Atlantic salmon stock from the escapees potentially hybridizing with it. In addition, SNPs have been shown to be the marker of choice for discrimination of different strains and species. Van Bers et al. (2012) [108] used a reduced representation library (RRL) to discover 3,569 SNPs for SNP assay development. Subsequently, a SNP assay with 384 loci was developed for tests of discrimination of individuals from different strains and species of tilapia.

In the realm of conservation genetics, genetic purity and hybridization identification using molecular genetic tools play an essential role in the conservation and management of morphologically similar fish strain and species [109]. For example, black bass species are among the most popular freshwater sport fishes in North America and have been widely translocated and introduced as part of large-scale stocking effort [110]. However, the weak genetic barriers of black basses make them prone to introgressive hybridization and genetic swamping [111]. Li et al. (2015) [45] discovered thousands of SNPs from the RNA-seq data and created an SNP panel with 25 SNPs in order to access the genetic integrity and hybridization in hatchery and wild populations of Florida bass (*Micropterus floridanus*) and largemouth bass (*M. salmoides*). Wilawan et al. (2019)

[111] developed similar SNP resources for other black bass species through GBS, followed by validation in additional samples using two panels of 64 SNPs. Further genotyping results from > 1300 bass indicated that the developed panels could robustly and clearly delineate fifteen black bass species and their hybrids. These SNP resources and panels are flexible and cost-effective tools that could augment ongoing efforts toward fish conservation and management.

5.2 Parentage Analysis

The analysis of parentage is a key facet of aquaculture and molecular ecology. Effective methods of parentage assignment can facilitate selective breeding, aid in inbreeding avoidance, and increase aquaculture production efficiency [112]. In addition, an understanding of parentage patterns can provide information for the study of sexual selection [113], conservation biology [113], and animal speciation [114]. For parentage assignment, SNPs are rapidly replacing microsatellites because of their high abundance across the genome, amenability for high throughput genotyping, ease of automation and scoring, and lower genotyping error rates [115]. SNP markers for parentage assignment have been developed in a range of aquatic species, including Pacific oyster (*Crassostrea gigas*) [116, 117], black tiger shrimp (*Penaeus monodon*) [118], blue mussel (*Mytilus galloprovincialis*) [119], common carp (*Cyprinus carpio*) [120], rainbow trout (*Oncorhynchus mykiss*) [112], sockeye salmon (*Oncorhynchus nerka*) [121], Delta Smelt (*Hypomesus transpacificus*) [122], and steelhead (*Oncorhynchus mykiss*) [123, 124]. Given the developments on the NGS front, as well as continued progress in analytical approaches, SNPs are now considered as the best marker type for parentage analysis.

5.3 Marker-Assisted Selection and Genomic Selection

Genomics technologies such as marker-assisted selection (MAS) or genomic selection (GS) are now extensively used for genetic enhancement of aquaculture species [125]. These technologies assume that markers associate at high frequency with the gene or quantitative trait locus (QTL) of interest due to genetic linkage. Selection based on these trait-associated markers can facilitate efficient and precise genetic enhancement programs [4]. Many performance and production traits are complex and quantitative. Therefore, the core step of MAS is to correlate genetic and phenotypic variation through procedures like QTL mapping and genome-wide association studies (GWAS) [4]. Progress of MAS/GS analyses has been greatly accelerated by the application of SNP arrays. SNP arrays have been developed in many aquaculture species, including Atlantic salmon (*Salmo salar*) [126-128], catfish (*Ictalurus punctatus* and *I. furcatus*) [46, 97], common carp *(Cyprinus carpio*) [129], European oyster (*Crassostrea gigas* and *Ostrea edulis*) [130], Pacific oyster (*Crassostrea gigas*) [96, 130, 131], Pacific-white shrimp (*Litopenaeus vannamei*) [132], rainbow trout (*Oncorhynchus mykiss*) [133], and silver-lipped pearl oyster (*Pinctada maxima*) [134]. These high-density SNP genotyping arrays provide robust data for downstream QTL or GWAS analysis, some of the best examples including growth improvement in oyster [135, 136], disease resistance in catfish [137, 138], and sex determination in salmon species [139-141]. Although MAS and GS provide powerful tools for the genetic enhancement of aquaculture species, the application of these technologies is still limited to just a few species due to the lack of genomic resources (e.g., SNPs, genetic maps. annotated genome information, and high-throughput genotyping platforms) [4]. As the efficiency of techniques for SNP discovery, genotyping and other genetic procedures improve, the opportunities to incorporate MAS or GS into breeding programs for aquaculture species will surely increase [142].

5.4 Population Genetic Analysis

Population genetics is the study of genetic variation between and within populations, as well as the examination of gene or allele changes over space and time. SNP markers are ideally suited for population genetic analyses as they represent the most widespread type of sequence variation in the genome [143]. SNPs located in the neutral genomic regions can provide a picture of genome-wide effects of neutral evolutionary forces, whereas non-neutral loci can be used to search for signatures of selection [143]. The measurement of genetic diversity is a key component of population genetics. In aquaculture, genetic evaluation of the broodstock is a useful asset for future genetic gain when establishing base populations for breeding programs [144]. For example, Rengmark et al. (2006) [145] used 26 SNPs to examine the average observed heterozygosity (a genetic diversity index) in salmon populations sourced from five rivers and two farmed stocks. After marker genotyping, they found that one of the farmed stocks contained the highest degree of genetic diversity and can be could for further selective breeding programs. Using a 57K SNP array that was initially developed for rainbow trout, Zhang et al. (2018) [146] recently genotyped seven aquaculture salmonid populations in China and demonstrated the SNP assay was applicable as a universal tool for population diversity evaluation among diverse salmonid species.

SNPs are also increasingly being used as molecular ecology tools for the study of population genetics in aquatic species. Genome-wide SNP data were used to infer phylogeographic history and tested the hypothesis of recent stickleback introduction into central Oregon [147]. Using the same data, the authors conducted a genome-wide analysis of genetic diversity (π) and Wright's inbreeding coefficient (F_{is}) to confirm introgressive hybridization among oceanic and freshwater stickleback populations. Another population genetic study has revealed population structure and high levels of interspecific gene flow within the recent radiation of *Alcolapia* cichlid fish using a high-density of SNP data, which provide some insights into the mechanisms generating biological diversity [148]. A large number of 440,817 SNPs were previously generated for Atlantic herring for the population genetic studies of adaptation and natural selection [149]. A similar case study was conducted on four overfished populations of Atlantic cod, and a set of 77 SNP loci were identified to be associated with population differentiation and local adaptation.

5.5 Sex Determination

Sex-determination mechanisms in aquatic species are remarkably diverse and complicated because they are often affected by interactions among genetic and environmental factors [150]. Sex-determination systems have attracted considerable attention due to their outsized implications in both theoretical and applied research [151]. For example, mono-sex populations are highly prized in commercial aquaculture and sport fisheries due to sexual growth dimorphism in fish [152]. Production of mono-sex stocks has been applied in several important aquaculture species including Nile tilapia (*Oreochromis niloticus*) [153, 154], Atlantic halibut (*Hippoglossus hippoglossus*) [155, 156], rainbow trout (*Oncorhynchus mykiss*) [157], European seabass (*Dicentrarchus labrax*) [158], and bluegill (*Lepomis macrochirus*) [159]. Mono-sex fish stocks can also be utilized as biological control agents for the conservation management of nonnative fish populations. Recently, brook trout super-males (with YY genotype) have been produced to drive exotic brook trout populations toward extirpation using a suite of sex markers (microsatellites and SNPs) and juvenile sex reversal methods [160].

Advances in NGS and RAD-seq or GBS techniques have made it possible to construct high-density genetic linkage maps and associate SNP markers with putative sex-determining QTLs. For example, a study by Palaiokostas et al. (2013) [155] has constructed a high-density linkage map for Atlantic halibut and identified an assay of 10 SNPs that are significantly associated with phenotypic sex. A similar study was conducted on European sea bass by genotyping a single full-sib family using RAD-seq [161]. In this study, Palaiokostas et al. (2015) built a high-density linkage map with 6,706 SNPs and found that sex-determining QTLs were distributed on four different chromosomes, indicating polygenic sex determination in sea bass. The practical application of sex-linked SNPs is possible through the use of SNP assays. For instance, sex-linked SNP markers in Chinook salmon have been successfully developed into a TaqMan assay and applied for phenotypic sex prediction [162]. Larson et al. (2016) [163] discovered seven sex-associated loci in sockeye salmon and developed an assay with two loci for the genotyping of salmon samples collected throughout North America. By combining these two SNP loci with a known sex-determining gene, *sdY*, they obtained a high assignment concordance (~90% accuracy) between genetic and phenotypic sex. Although these SNP assays were found to be highly effective in terms of aquaculture management and production, the genotype prediction and phenotypic sex is not always perfectly matched [162, 163]. Future research is necessary to fully understand the mechanisms underlying sex-determination in aquatic species.

6. Further SNP Development and Application in Aquaculture and Conservation Studies

Although reduced representation approaches such as transcriptome (RNA-seq) and restriction site-associated nuclear DNA sequencing have become increasingly central to SNP development in aquaculture and conservation fish species, some weaknesses still exist in these methods. For example, RNA-seq-based SNP discovery targets the loci restricted to the genic regions, which only encompasses a small portion of the genome-wide SNPs [102]. The restriction site-associated nuclear DNA sequencing data, however, are less robust when RE sites include *de* *novo* mutation, and/or fragmented DNA is used (i.e., allele dropout) [164-166]. On the other hand, amplicon sequencing and sequencing capture methods such as Genotyping-in-Thousands by sequencing (GT-seq) [90], Multiplexed PCR Targeted Amplicon sequencing (MTA-seq) [167], and Highly Multiplexed Amplicon sequencing (HiMAP) [168], have been developed further. These approaches take advantage of PCR and only target the SNPs with prior sequence information, leading to high repeatability of the genotyping data. With these approaches, researchers can maximize the balance between cost (can be low as $6 per sample in GT-seq, [169]), sample size, and the number of SNPs (up to 6,144 loci with AmpliSeq technology, [164]). A recent case study using amplicon sequencing technique designed a panel targeting 3,187 SNPs in fugu (*Takifugu rubripes*) and applied it for genotyping of 652 individuals across 10 full-sib families [164]. Consequently, a total of 2,655 SNPs was retained after data filtering and successfully applied for population structure and sex determination analyses among examined fugu populations. These results highlight the potential of amplicon sequencing in conservation genetic studies, as well as molecular breeding for improved aquaculture management and production.

Study Species Background Information

7. Florida Bass Parentage Assignment

The Florida bass (*Micropterus floridanus*) is a highly prized sportfish believed to be natively restricted to peninsular Florida; however, its current distribution is international [170]. Florida bass is a member of the black basses (genus *Micropterus*), which is one of the most popular game fish groups in the United States. Evidence has shown that the ancestral *Micropterus* representative began allopatric speciation about 10 million years ago, and up to the present, there are 14 species recognized in this genus [171]. The Florida bass (*M. floridanus*) and the Northern largemouth bass

(*M. salmoides*) are collectively known as largemouth bass but were separated into two subspecies in 1949 by Bailey and Hubbs. It is estimated that the Florida bass diverged from the Northern largemouth bass less than 5 million years ago [171]. The distribution of Florida bass overlaps with that of the Northern largemouth bass forming a natural hybrid zone in the southeastern US; however, the scope of this introgression has been expanded dramatically via stocking efforts [45, 110, 172, 173]. The Florida bass is important for its economic value in fisheries and sport fishing due to its tendency to attain a larger maximum size and aggressiveness relative to Northern largemouth bass. It was estimated that Florida bass fishing contributes $632 million per year to the economy of Florida (U.S. Department of Interior, 2006, [174]). Supplemental stocking of Florida bass is a common management strategy for augmenting natural production in low-recruitment water systems [175]. Additional forms of stocking involve the transplanting of Florida bass to non-native systems for the specific fisheries attributes enhancement (e.g. growth) [176].

Parentage analyses are critical in tracking ecological or life-history characteristics of released fish, estimating genetic parameters and breeding values, and minimizing inbreeding in broodstock [112, 177]. Previous parentage assignment of Florida bass has been conducted with different molecular markers such as allozymes and microsatellites [13, 170, 178-181]. Philipp (1991) [13] utilized a set of three allozyme loci for parentage analysis in Florida bass, Northern largemouth bass, and their F1 hybrids, and evaluated their reproductive success in experimental populations. A shift toward DNA-based marker technology for Florida bass parentage analysis was seen in the 2000s as molecular technology improved. Lutz-Carrillo et al. optimized 11 (2006) [180] and 52 (2008) [170] microsatellite loci for Florida bass and other micropterids, and demonstrated that these genetic resources are appropriate for studies ranging from parentage analyses to taxon identification. A similar study was conducted by Seyoum et al. (2013) [181], who isolated and characterized an additional 18 microsatellites for the largemouth bass. With these microsatellite marker resources, Austin et al. (2012) [178] screened broodstock and offspring in a raceway breeding design to assess hatchery effects on genetic diversity and to determine sibling relationships. While their assessment suggests that raceway breeding practices were sufficient to maintain genetic diversity, they found that a set of nine microsatellites were not able to assign all individual fry due to low heterozygosity of the data set. A subsequent study by Hargrove and Austin (2017) [179] utilized a different set of microsatellites for parentage analysis and investigated the mating patterns of the Florida bass spawning in raceways. Nevertheless, the assignment power and the minimum number of microsatellite markers needed to accurately conduct parentage analysis in Florida bass were not comprehensively estimated. In addition, SNP markers that are sufficient for parentage analysis in the Florida bass are not yet available, but this scarcity will be covered in this dissertation.

8. White Bass Population Genetics

The white bass (*Morone chrysops*) is a temperate bass species with high commercial and ecological value in North America [182]. The white bass was historically found only in the Great Lakes and Mississippi River drainages [183]. However, it has been widely introduced outside its native range for stocking and food production purposes. In terms of an ecological role, white bass are important as intermediate predators in ecosystems. Meanwhile, they are food sources for larger fish and other predators. In addition to their ecosystem services, white bass are also important for their economic value in fisheries and aquaculture. The hybrid striped bass (HSB) is a cross-breeding aquaculture product resulting from the crossing of the two parent species, white bass and the striped bass (*M. chrysops* × *M. saxatilis*) [184]. The HSB is prized because of its improved performance in growth, survival, hardiness, and disease resistance, presumably due to hybrid vigor (or heterosis) [184]. Commercial production of HSB began in the early 1980s with the original HSB or palmetto bass (striped bass female × white bass male), while current HSB has been mostly replaced by the more easily spawned reciprocal cross or sunshine bass (striped bass male × white bass female) [185]. Although HSB is a valuable sector of US aquaculture production (valued at nearly $30 million in 2012, [184-187]), the high market prices and increases in input costs of production have hampered the growth of the HSB industry [188]. It is believed that a selective breeding program aimed at reducing production costs should be an effective way to promote the HSB farming industry. To establish a successful breeding program, basic information on the biology and genetics of the parent species of the hybrid is needed. A domestication program for striped and white bass was initiated during the 1980s - 1990s, with its goal of acclimating wild fish to culture conditions and evaluating production traits [189-191]. Currently, the US HSB industry is sourcing wild striped and white bass broodstock from river systems across the US, aiming at providing the industry with a superior and sustainable, cultured broodstock source [189, 192, 193].

Previous assessments of population genetics of striped bass have been conducted in an area ranging from the Miramichi River and the Shubenacadie River system in Canada [194] to the Gulf of Mexico and Atlantic coast [195] using several molecular markers including mtDNA, RAPD, RFLP, microsatellites, and SNPs [193-207]. For example, Chapman (1990) [208] used mtDNA to examine genetic variation in upper and lower Chesapeake Bay striped bass populations. In this study, they observed genetically distinct populations in the Chesapeake Bay and hypothesized that migration patterns and sexual variation in homing response might lead to distinct genetic patterns among examined populations. However, Laughlin and Turner (1996) [209] reported very little genetic differentiation for multilocus variable number of tandem repeat (VNTR) markers among striped bass populations in the Chesapeake Bay. It remains uncertain if the multiple tributaries within Chesapeake Bay constitute distinct management units of striped bass. Subsequently, Brown et al. (2005) [207] reconciled the conflicting results and demonstrated a single Chesapeake Bay management unit using 10 microsatellites. A low level of population differentiation was observed in other striped bass populations. Using RFLP and three nuclear loci, Diaz et al. (1997) [210] reported low levels of genetic divergence among the Atlantic and Gulf of Mexico populations of striped bass. The following study by Roy et al. (2000) [195] also found a similar genetic pattern of genetic differentiation between the Atlantic and Gulf of Mexico populations using eight microsatellites. Similar low levels of genetic polymorphism were also reported in white bass. White (2000) [211] used protein electrophoresis at seven loci to examine genetic variation in Ohio River drainage white bass populations, noting exceptionally low heterozygosity and concluding that he could not detect ‘genetically meaningful stock structure outside of Lake Erie.’ More recently, Couch et al. (2006) [212] identified 149 novel microsatellites from striped and white bass. However, only six white bass individuals were tested using these markers and the authors noted an average of only 2.2 alleles per locus. Given the low level of genetic polymorphism in *Morone* species, future research using highly polymorphic genetic markers on samples collected throughout the native range and from domesticated broodstocks may shed light on the fine-scale assessment of genetic structure and the potential for selective breeding improvement.

9. Walleye Population Genetics

The walleye (*Sander vitreus*) is an ecologically and economically important fish species that belongs to the family Percidae [213]. They are found in the freshwater river and lake systems throughout North America [214]. The walleye is a cool-water fish species with optimal growth temperatures ranging from 18-22 ^{o}C [215]. The native distribution of walleye spans a latitudinal range that extends northward to Mackenzie River at the Arctic coast and southward to the Mobile River Basin in the southern US [216]. Walleye populations have been widely introduced outside their native range to support large sport and commercial fisheries [217], particularly in the western watersheds (e.g., Columbia River system, western Montana reservoirs, Missouri River drainage [218]). Meanwhile, stocking to supplement the existing populations, or restocking, is a widespread approach for walleye management and conservation [219]. Both introduction and restocking practices should take into account the population structure and genetic diversity of the species [220]. Given their importance in ecological functions and recreational angling, there is a requirement for reliable molecular markers for population genetics studies in walleye.

Previous population genetics studies in walleye have been conducted using several molecular markers including allozymes, mtDNA, RFLP, microsatellites, and SNPs [213, 221-241]. Early genetic surveys using allozymes generally failed to reveal substantial genetic differentiation among walleye populations [241, 242]. For example, Ward et al. (1989) [241] reported evidence of geographic patterning of allele frequencies in 15 population of walleye from the Great Lakes and northern Manitoba using nine allozymes. However, the genetic variation among the same walleye populations was three to five times greater when mtDNA haplotypes were used for data analysis, suggesting the relatively low abundance and low level of polymorphism of allozyme markers [243]. This study also revealed the existence of three distinct genetic groups among the walleye populations: the Missouri refugium, the Mississippi refugium, and the Atlantic refugium [241]. The same genetic patterns were additionally confirmed by population genetic studies using mtDNA markers [236, 244], suggesting the long-term isolation and lack of gene flow among these walleye groups. A fourth walleye group with unique mtDNA haplotypes was identified from the unglaciated Eastern Highlands regions (Ohio, Kentucky, and Tennessee River Basins) [236, 239]. From mtDNA analysis, Billington et al. (1992) [235] noted an additional walleye group from the Mobile River Basin that possessed a unique haplotype (haplotype 34). This unique genetic pattern was later proven to be highly divergent from other North America walleye populations using mtDNA analysis [244, 245]. Later population structure and genetic divergence among walleye populations were characterized by microsatellite loci. Stepien et al. (2009) [230] first utilized ten microsatellites to analyze the genetic structure of walleye across its native range and compared the broad- vs. fine-scale divergence pattern among 921 walleye individuals from 26 spawning sites. The microsatellite-based data analysis highlighted the low genetic diversity and the genetic distinctiveness of walleye from the Mobile River Basin, as well as the unique genetic patterns of walleye populations from the Ohio River system. Similar population patterns were reported by Haponski and Stepien (2014) [233] using nine microsatellites and mtDNA control regions sequences. Taken together, they hypothesized that these population structures and genetic patterns were shaped by climate change and drainage connections, with northern walleye patterns attributable to post-glacial recolonization [233].

Recently, SNP markers for walleye were identified from restriction site-associated DNA sequencing and GT-seq techniques [169, 222, 246]. Chen (2016) identified a total of 12,264 SNPs from RAD-seq data and applied these SNP loci for population assignment and quantification of natal philopatry in Lake Erie walleye. Allen et al. (2017) [222] characterized 1,081 SNPs from GBS data and used these loci to measure the genetic variation of walleye from several lakes and watersheds in Alberta, Canada. A recent review paper by Meek and Larson (2019) [169] indicated that multiple sequence capture panels containing 500-10,000 loci were developed for genetic management and population structure and diversity investigation across walleye populations in the

Great Lakes regions, while the practical application of these panels has not been reported to date. The studies highlighted above, however, focused their efforts on SNP marker development and utilization in northern walleye populations, limiting their potential applicability to more restricted southern walleye populations (the focus of the final chapter of my dissertation).

10. Dissertation Overview

Given that SNPs are increasingly utilized for improved selection and management in aquaculture and conservation fish species (as reviewed above), I developed genomic SNP resources and validated, multiplexed SNP panels in three key aquatic species in the southeast US. **Chapter II** focuses on the development of SNP marker panels for use in parentage analysis of the Florida bass. The studies include the identification of SNPs from GBS, marker validation, the development of multiplexed SNPs for use in the MassARRAY system, a comparison of parentage accuracy between microsatellites and SNPs, and an examination of the assignment power using smaller subsets of the developed SNP markers. In **Chapter III**, I employ GBS techniques to identify informative SNP markers for the use in population genetic analyses in white bass. The studies include the identification of SNPs from GBS, marker validation, the development of multiplexed SNP panels, and an examination of the SNP panels in assigning white bass individuals back to their origin populations. In **Chapter IV**, I describe the genome and SNP resource development for use in walleye population identification and hybridization analysis. The studies include the identification of SNPs from GBS, the development of multiplexed SNP panels, and subsequent utilization of these SNPs in rapid and accurate identification of genetic purity and classification of various (northern/southern) hybrid classes among walleye individuals. These SNP resources represent a valuable tool for a myriad of downstream applications in genetic management, stock enhancement, and population management of these fish species.

References

1. Liu, Z.J., Bioinformatics in Aquaculture: Principles and Methods. 2017: John Wiley & Sons.
2. Beck, B.H. and E. Peatman, Mucosal health in aquaculture. 2015: Academic Press.
3. Yáñez, J.M., S. Newman, and R.D. Houston, Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front Genet, 2015. **6**: p. 128.
4. Abdelrahman, H., et al., Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. BMC Genomics, 2017. **18**(1): p. 191.
5. Allendorf, F.W., P.A. Hohenlohe, and G. Luikart, Genomics and the future of conservation genetics. Nature Reviews Genetics, 2010. **11**(10): p. 697.
6. Liu, Z.J. and J.F. Cordes, DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 2004. **238**(1-4): p. 1-37.
7. Johnson, K., J. Wright, and B. May, Linkage relationships reflecting ancestral tetraploidy in salmonid fish. Genetics, 1987. **116**(4): p. 579-591.
8. Liu, Q., et al., Gene-centromere mapping of six enzyme loci in gynogenetic channel catfish. Journal of Heredity, 1992. **83**(4): p. 245-248.
9. Seeb, J.E. and L.W. Seeb, Gene mapping of isozyme loci in chum salmon. Journal of Heredity, 1986. **77**(6): p. 399-402.
10. Buroker, N.E., Population genetics of the American oyster *Crassostrea virginica* along the Atlantic coast and the Gulf of Mexico. Marine Biology, 1983. **75**(1): p. 99-112.

11. Dunham, J.B. and W. Minckley, Allozymic variation in desert pupfish from natural and artificial habitats: genetic conservation in fluctuating populations. Biological Conservation, 1998. **84**(1): p. 7-15.

12. King, T.L., R. Ward, and E.G. Zimmerman, Population structure of eastern oysters (*Crassostrea virginica*) inhabiting the Laguna Madre, Texas, and adjacent bay systems. Canadian Journal of Fisheries and Aquatic Sciences, 1994. **51**(S1): p. 215-222.

13. Philipp, D.P., Genetic implications of introducing Florida largemouth bass, *Micropterus salmoides floridanus*. Canadian Journal of Fisheries and Aquatic Sciences, 1991. **48**(S1): p. 58-65.

14. English, L., et al., Allozyme variation in three generations of selection for whole weight in Sydney rock oysters (*Saccostrea glomerata*). Aquaculture, 2001. **193**(3-4): p. 213-225.

15. McGoldrick, D.J. and D. Hedgecock, Fixation, segregation and linkage of allozyme loci in inbred families of the Pacific oyster *Crassostrea gigas* (Thunberg): implications for the causes of inbreeding depression. Genetics, 1997. **146**(1): p. 321-334.

16. McAndrew, B.J. and K.C. Majumdar, Tilapia stock identification using electrophoretic markers. Aquaculture, 1983. **30**(1-4): p. 249-261.

17. Botstein, D., et al., Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics, 1980. **32**(3): p. 314.

18. Funkenstein, B., et al., Restriction site polymorphism of mitochondrial DNA of the gilthead sea bream (*Sparus aurata*) broodstock in Eilat, Israel. Aquaculture, 1990. **89**(3-4): p. 217-223.

19. Karl, S.A. and J.C. Avise, Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science, 1992. **256**(5053): p. 100-102.

20. Russell, V.J., et al., Use of restriction fragment length polymorphism to distinguish between salmon species. Journal of Agricultural and Food Chemistry, 2000. **48**(6): p. 2184-2188.

21. Williams, J.G., et al., DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research, 1990. **18**(22): p. 6531-6535.

22. Welsh, J. and M. McClelland, Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research, 1990. **18**(24): p. 7213-7218.

23. Wirgin, I.I. and J.R. Waldman, What DNA can do for you. Fisheries, 1994. **19**(7): p. 16-27.

24. Vos, P., et al., AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, 1995. **23**(21): p. 4407-4414.

25. Bensch, S. and M. Åkesson, Ten years of AFLP in ecology and evolution: why so few animals? Molecular Ecology, 2005. **14**(10): p. 2899-2914.

26. Litt, M. and J.A. Luty, A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. American Journal of Human Genetics, 1989. **44**(3): p. 397.

27. Tautz, D., Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research, 1989. **17**(16): p. 6463-6471.

28. Abdul-Muneer, P., Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. Genetics Research International, 2014. **2014**.

29. Garvin, M., K. Saitoh, and A. Gharrett, Application of single nucleotide polymorphisms to non-model species: a technical review. Molecular Ecology Resources, 2010. **10**(6): p. 915-934.

30. Morin, P.A., G. Luikart, and R.K. Wayne, SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution, 2004. **19**(4): p. 208-216.

31. Group, I.S.M.W., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 2001. **409**(6822): p. 928.

32. Slate, J., et al., Gene mapping in the wild with SNPs: guidelines and future directions. Genetica, 2009. **136**(1): p. 97-107.

33. Morin, P.A., et al., Characterization of 18 SNP markers for sperm whale (*Physeter macrocephalus*). Molecular Ecology Notes, 2007. **7**(4): p. 626-630.

34. Smith, C., et al., Characterization of 13 single nucleotide polymorphism markers for chum salmon. Molecular Ecology Notes, 2005. **5**(2): p. 259-262.

35. Elfstrom, C., et al., Characterization of 12 single nucleotide polymorphisms in weathervane scallop. Molecular Ecology Notes, 2005. **5**(2): p. 406-409.

36. Adams, M.D., et al., Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 1991. **252**(5013): p. 1651-1656.

37. Moen, T., et al., A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. BMC Genomics, 2008. **9**(1): p. 223.

38. Hayes, B., et al., An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. Aquaculture, 2007. **265**(1-4): p. 82-90.

39. Wang, S., et al., Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics, 2008. **9**(1): p. 450.

40. Zhu, C., et al., Development and characterization of new single nucleotide polymorphism markers from expressed sequence tags in common carp *(Cyprinus carpio*). International Journal of Molecular Sciences, 2012. **13**(6): p. 7343-7353.

41. Zhang, M., et al., Isolation and characterization of 25 novel EST-SNP markers in grass carp (*Ctenopharyngodon idella*). Conservation Genetics Resources, 2015. **7**(4): p. 819-822.

42. Kim, J.E., et al., Development and validation of single nucleotide polymorphism (SNP) markers from an expressed sequence tag (EST) database in olive flounder (*Paralichthys olivaceus*). Development & Reproduction, 2014. **18**(4): p. 275.

43. Vera, M., et al., Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, *Scophthalmus maximus*, database. Aquaculture, 2011. **313**(1-4): p. 31-41.

44. Kong, L., J. Bai, and Q. Li, Comparative assessment of genomic SSR, EST–SSR and EST–SNP markers for evaluation of the genetic diversity of wild and cultured Pacific oyster, *Crassostrea gigas* Thunberg. Aquaculture, 2014. **420**: p. S85-S91.

45. Li, C., et al., Discovery and validation of gene-linked diagnostic SNP markers for assessing hybridization between Largemouth bass (*Micropterus salmoides*) and Florida bass (*M. floridanus*). Molecular Ecology Resources, 2015. **15**(2): p. 395-404.

46. Liu, S., et al., Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics, 2011. **12**(1): p. 53.

47. Baird, N.A., et al., Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One, 2008. **3**(10): p. e3376.

48. Elshire, R.J., et al., A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One, 2011. **6**(5): p. e19379.

49. Wang, S., et al., 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature Methods, 2012. **9**(8): p. 808.

50. Peterson, B.K., et al., Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One, 2012. **7**(5): p. e37135.

51. Toonen, R.J., et al., ezRAD: a simplified method for genomic genotyping in non-model organisms. PeerJ, 2013. **1**: p. e203.

52. Collins, L.J., et al., An approach to transcriptome analysis of non-model organisms using short-read sequences, in Genome Informatics 2008: Genome Informatics Series Vol. 21. 2008, World Scientific. p. 3-14.

53. Vera, J.C., et al., Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology, 2008. **17**(7): p. 1636-1647.

54. Barbazuk, W.B., et al., SNP discovery via 454 transcriptome sequencing. The Plant Journal, 2007. **51**(5): p. 910-918.

55. Zhao, Y., et al., A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics, 2019. **20**(1): p. 160.

56. Miller, N.A., et al., Management of high-throughput DNA sequencing projects: Alpheus. Journal of Computer Science and Systems Biology, 2008. **1**: p. 132.

57. Grabherr, M.G., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 2011. **29**(7): p. 644.

58. Chevreux, B., et al., Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research, 2004. **14**(6): p. 1147-1159.

59. Simpson, J.T., et al., ABySS: a parallel assembler for short read sequence data. Genome Research, 2009. **19**(6): p. 1117-1123.

60. Zerbino, D.R. and E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 2008. **18**(5): p. 821-829.

61. Schulz, M.H., et al., Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics, 2012. **28**(8): p. 1086-1092.

62. Ning, Z., A.J. Cox, and J.C. Mullikin, SSAHA: a fast search method for large DNA databases. Genome Research, 2001. **11**(10): p. 1725-1729.

63. Kent, W.J., BLAT-the BLAST-like alignment tool. Genome Research, 2002. **12**(4): p. 656-664.

64. Li, H., J. Ruan, and R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 2008. **18**(11): p. 1851-1858.

65. Jiang, H. and W.H. Wong, SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics, 2008. **24**(20): p. 2395-2396.

66. Li, R., et al., SOAP: short oligonucleotide alignment program. Bioinformatics, 2008. **24**(5): p. 713-714.

67. Langmead, B., et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 2009. **10**(3): p. R25.

68. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009. **25**(14): p. 1754-1760.

69. Li, H., et al., The sequence alignment/map format and SAMtools. Bioinformatics, 2009. **25**(16): p. 2078-2079.

70. Quinlan, A.R. and I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010. **26**(6): p. 841-842.

71. DePristo, M.A., et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 2011. **43**(5): p. 491.

72. Garrison, E. and G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907, 2012.

73. Baranski, M., et al., The development of a high density linkage map for black tiger shrimp (*Penaeus monodon*) based on cSNPs. PLoS One, 2014. **9**(1): p. e85413.

74. Salem, M., et al., RNA-Seq identifies SNP markers for growth traits in rainbow trout. PLoS One, 2012. **7**(5): p. e36264.

75. Davey, J.W., et al., Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 2011. **12**(7): p. 499.

76. Robledo, D., et al., Applications of genotyping by sequencing in aquaculture breeding and genetics. Reviews in Aquaculture, 2018. **10**(3): p. 670-682.

77. Erlich, Y., et al., DNA Sudoku-harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Research, 2009. **19**(7): p. 1243-1253.

78. Meyer, M., U. Stenzel, and M. Hofreiter, Parallel tagged sequencing on the 454 platform. Nature Protocols, 2008. **3**(2): p. 267.

79. Meyer, M., et al., Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Research, 2007. **35**(15): p. e97.

80. Andrews, K.R., et al., Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 2016. **17**(2): p. 81.

81. Bradbury, P.J., et al., TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics, 2007. **23**(19): p. 2633-2635.

82. Hohenlohe, P.A., et al., Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS genetics, 2010. **6**(2): p. e1000862.

83. Nielsen, R., et al., SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One, 2012. **7**(7): p. e37558.

84. Catchen, J., et al., Stacks: an analysis tool set for population genomics. Molecular Ecology, 2013. **22**(11): p. 3124-3140.

85. Eaton, D.A., PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. Bioinformatics, 2014. **30**(13): p. 1844-1849.

86. Shendure, J. and H. Ji, Next-generation DNA sequencing. Nature Biotechnology, 2008. **26**(10): p. 1135.

87. Hohenlohe, P.A., et al., Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired‐end RAD sequencing. Molecular Ecology, 2013. **22**(11): p. 3002-3013.

88. Schweyen, H., A. Rozenberg, and F. Leese, Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. The Biological Bulletin, 2014. **227**(2): p. 146-160.

89. Andrews, K.R., et al., Trade‐offs and utility of alternative RADseq methods: Reply to Puritz et al. Molecular Ecology, 2014. **23**(24): p. 5943-5946.

90. Campbell, N.R., S.A. Harmon, and S.R. Narum, Genotyping‐in‐Thousands by sequencing (GT‐seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Eesources, 2015. **15**(4): p. 855-867.

91. Seeb, J., et al., Single‐nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Eesources, 2011. **11**: p. 1-8.

92. Wang, J., et al., High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. BMC Genomics, 2009. **10**(1): p. 561.

93. Spurgeon, S.L., R.C. Jones, and R. Ramakrishnan, High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. PLoS One, 2008. **3**(2): p. e1662.

94. Campbell, N., et al., Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa. Molecular Ecology Eesources, 2012. **12**(5): p. 942-949.

95. Schunter, C., et al., Kinship analyses identify fish dispersal events on a temperate coastline. Proceedings of the Royal Society B: Biological Sciences, 2014. **281**(1785): p. 20140556.

96. Qi, H., et al., Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*). PLoS One, 2017. **12**(3): p. e0174007.

97. Zeng, Q., et al., Development of a 690 K SNP array in catfish and its application for genetic mapping and validation of the reference genome sequence. Scientific Reports, 2017. **7**: p. 40347.

98. Liu, S., et al., Development of the catfish 250K SNP array for genome-wide association studies. BMC Research Notes, 2014. **7**(1): p. 135.

99. Gabriel, S., L. Ziaugra, and D. Tabbaa, SNP genotyping using the Sequenom MassARRAY iPLEX platform. Current Protocols in Human Genetics, 2009. **60**(1): p. 2.12. 11-12.12. 18.

100. Hargrove, J.S., et al., Using species-diagnostic SNPs to detail the distribution and dynamics of hybridized black bass populations in southern Africa. Biological Invasions, 2019. **21**(5): p. 1499-1509.

101. Zhao, H., et al., SNP marker panels for parentage assignment and traceability in the Florida bass (*Micropterus floridanus*). Aquaculture, 2018. **485**: p. 30-38.

102. Li, C., et al., SNP discovery in wild and domesticated populations of blue catfish, I ctalurus furcatus, using genotyping - by - sequencing and subsequent SNP validation. Molecular Ecology Eesources, 2014. **14**(6): p. 1261-1270.

103. Jung, H., et al., A candidate gene association study for growth performance in an improved giant freshwater prawn (*Macrobrachium rosenbergii*) culture line. Marine Biotechnology, 2014. **16**(2): p. 161-180.

104. Malde, K., et al., Whole genome resequencing reveals diagnostic markers for investigating global migration and hybridization between minke whale species. BMC Genomics, 2017. **18**(1): p. 76.

105. Krück, N.C., D.I. Innes, and J.R. Ovenden, New SNP s for population genetic analysis reveal possible cryptic speciation of eastern Australian sea mullet (*Mugil cephalus*). Molecular Ecology Eesources, 2013. **13**(4): p. 715-725.

106. Karlsson, S., et al., Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP‐chip. Molecular Ecology Eesources, 2011. **11**: p. 247-253.

107. Pritchard, V.L., et al., Single nucleotide polymorphisms to discriminate different classes of hybrid between wild Atlantic salmon and aquaculture escapees. Evolutionary Applications, 2016. **9**(8): p. 1017-1031.

108. Van Bers, N., et al., SNP marker detection and genotyping in tilapia. Molecular Ecology Eesources, 2012. **12**(5): p. 932-941.

109. Bickford, D., et al., Cryptic species as a window on diversity and conservation. Trends in Ecology & Evolution, 2007. **22**(3): p. 148-155.

110. Barthel, B.L., et al., Genetic relationships among populations of Florida Bass. Transactions of the American Fisheries Society, 2010. **139**(6): p. 1615-1641.

111. Thongda, W., et al., Species-diagnostic SNP markers for the black basses (*Micropterus* spp.): a new tool for black bass conservation and management. Conservation Genetics Resources, 2019: p. 1-10.

112. Liu, S., et al., Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture, 2016. **452**: p. 178-182.

113. Serbezov, D., et al., Mating patterns and determinants of individual reproductive success in brown trout (*Salmo trutta*) revealed by parentage analysis of an entire stream living population. Molecular Ecology, 2010. **19**(15): p. 3193-3205.

114. Muhlfeld, C.C., et al., Hybridization rapidly reduces fitness of a native trout in the wild. Biology Letters, 2009. **5**(3): p. 328-331.

115. Flanagan, S.P. and A.G. Jones, The future of parentage analysis: From microsatellites to SNPs and beyond. Molecular Ecology, 2019. **28**(3): p. 544-567.

116. Lapègue, S., et al., Development of SNP-genotyping arrays in two shellfish species. Molecular Ecology Eesources, 2014. **14**(4): p. 820-830.

117. Jin, Y.L., et al., Development, inheritance and evaluation of 55 novel single nucleotide polymorphism markers for parentage assignment in the Pacific oyster (*Crassostrea gigas*). Genes & Genomics, 2014. **36**(2): p. 129-141.

118. Sellars, M.J., et al., Comparison of microsatellite and SNP DNA markers for pedigree assignment in B lack T iger shrimp, *Penaeus monodon.* Aquaculture Research, 2014. **45**(3): p. 417-426.

119. Nguyen, T.T., B.J. Hayes, and B.A. Ingram, Genetic parameters and response to selection in blue mussel (*Mytilus galloprovincialis*) using a SNP-based pedigree. Aquaculture, 2014. **420**: p. 295-301.

120. Xu, J., et al., Development and evaluation of a high‐throughput single nucleotide polymorphism multiplex assay for assigning pedigrees in common carp. Aquaculture Research, 2017. **48**(4): p. 1866-1876.

121. Hauser, L., et al., An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Molecular Ecology Eesources, 2011. **11**: p. 150-161.

122. Lew, R.M., et al., Using next-generation sequencing to assist a conservation hatchery: a single-nucleotide polymorphism panel for the genetic management of endangered delta smelt. Transactions of the American Fisheries Society, 2015. **144**(4): p. 767-779.

123. Steele, C.A., et al., A validation of parentage-based tagging using hatchery steelhead in the Snake River basin. Canadian Journal of Fisheries and Aquatic Sciences, 2013. **70**(7): p. 1046-1054.

124. Abadía‐Cardoso, A., et al., Large‐scale parentage analysis reveals reproductive patterns and heritability of spawn timing in a hatchery population of steelhead (*Oncorhynchus mykiss*). Molecular Ecology, 2013. **22**(18): p. 4733-4746.

125. Zenger, K.R., et al., Genomic selection in aquaculture: Application, limitations and opportunities with special reference to marine shrimp and pearl oysters. Frontiers in Genetics, 2018. **9**: p. 693.

126. Lien, S., et al., A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics, 2011. **12**(1): p. 615.

127. Yáñez, J.M., et al., Genomewide single nucleotide polymorphism discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. Molecular Ecology Resources, 2016. **16**(4): p. 1002-1011.

128. Houston, R.D., et al., Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar).* BMC Genomics, 2014. **15**(1): p. 90.

129. Xu, J., et al., Development and evaluation of the first high-throughput SNP array for common carp (*Cyprinus carpio*). BMC Genomics, 2014. **15**(1): p. 307.

130. Gutierrez, A.P., et al., Development of a medium density combined-species SNP array for Pacific and European oysters (*Crassostrea gigas* and *Ostrea edulis*). G3: Genes, Genomes, Genetics, 2017. **7**(7): p. 2209-2218.

131. Hedgecock, D., et al., Second-generation linkage maps for the pacific oyster Crassostrea gigas reveal errors in assembly of genome scaffolds. G3: Genes, Genomes, Genetics, 2015. **5**(10): p. 2007-2019.

132. Jones, D.B., et al., A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, *Litopenaeus vannamei.* Scientific Reports, 2017. **7**(1): p. 10360.

133. Palti, Y., et al., The development and characterization of a 57 K single nucleotide polymorphism array for rainbow trout. Molecular Ecology Resources, 2015. **15**(3): p. 662-672.

134. Jones, D.B., et al., Genome-wide SNP validation and mantle tissue transcriptome analysis in the silver-lipped pearl oyster, *Pinctada maxima.* Marine Biotechnology, 2013. **15**(6): p. 647-658.

135. Jones, D.B., et al., Determining genetic contributions to host oyster shell growth: quantitative trait loci and genetic association analysis for the silver-lipped pearl oyster, *Pinctada maxima.* Aquaculture, 2014. **434**: p. 367-375.

136. Guo, X., et al., Genetic mapping and QTL analysis of growth-related traits in the Pacific oyster. Marine Biotechnology, 2012. **14**(2): p. 218-226.

137. Zhou, T., et al., GWAS analysis of QTL for enteric septicemia of catfish and their involved genes suggest evolutionary conservation of a molecular mechanism of disease resistance. Molecular Genetics and Genomics, 2017. **292**(1): p. 231-242.

138. Geng, X., et al., A genome-wide association study in catfish reveals the presence of functional hubs of related genes within QTLs for columnaris disease resistance. BMC Genomics, 2015. **16**(1): p. 196.

139. Barson, N.J., et al., Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. Nature, 2015. **528**(7582): p. 405.

140. Ayllon, F., et al., The vgll3 locus controls age at maturity in wild and domesticated Atlantic salmon (*Salmo salar L.*) males. PLoS Genetics, 2015. **11**(11): p. e1005628.

141. Pedersen, S., et al., Quantitative trait loci for precocious parr maturation, early smoltification, and adult maturation in double-backcrossed trans-Atlantic salmon (*Salmo salar*). Aquaculture, 2013. **410**: p. 164-171.

142. Yue, G.H., Recent advances of genome mapping and marker‐assisted selection in aquaculture. Fish and Fisheries, 2014. **15**(3): p. 376-396.

143. Helyar, S., et al., Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources, 2011. **11**: p. 123-136.

144. Fernández, J., et al., Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress. Frontiers in Genetics, 2014. **5**: p. 414.

145. Rengmark, A.H., et al., Genetic variability in wild and farmed Atlantic salmon (*Salmo salar*) strains estimated by SNP and microsatellites. Aquaculture, 2006. **253**(1-4): p. 229-237.

146. Zhang, H.-Y., et al., Population genetic analysis of aquaculture salmonid populations in China using a 57K rainbow trout SNP array. PLoS One, 2018. **13**(8): p. e0202582.

147. Catchen, J., et al., The population structure and recent colonization history of O regon threespine stickleback determined using restriction ‐ site associated DNA ‐ sequencing. Molecular Ecology, 2013. **22**(11): p. 2864-2883.

148. Ford, A.G., et al., High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment. Molecular Ecology, 2015. **24**(13): p. 3421-3440.

149. Lamichhaney, S., et al., Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences, 2012. **109**(47): p. 19345-19350.

150. Shi, X., et al., Female-specific SNP markers provide insights into a WZ/ZZ sex determination system for mud crabs *Scylla paramamosain*, *S. tranquebarica* and *S. serrata* with a rapid method for genetic sex identification. BMC Genomics, 2018. **19**(1): p. 981.

151. Pennell, M.W., J.E. Mank, and C.L. Peichel, Transitions in sex determination and sex chromosomes across vertebrate species. Molecular Ecology, 2018. **27**(19): p. 3950-3963.

152. Dunham, R., Production and use of monosex or sterile fishes in aquaculture. Reviews in Aquatic Sciences, 1990. **2**(1): p. 1-17.

153. Chakraborty, S.B., et al., Growth of mixed-sex and monosex Nile tilapia in different culture systems. Turkish Journal of Fisheries and Aquatic Sciences, 2011. **11**(1): p. 131-138.

154. Dan, N.C. and D.C. Little, The culture performance of monosex and mixed-sex new-season and overwintered fry in three strains of Nile tilapia (*Oreochromis niloticus*) in northern Vietnam. Aquaculture, 2000. **184**(3-4): p. 221-231.

155. Palaiokostas, C., et al., Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. BMC Genomics, 2013. **14**(1): p. 566.

156. Hendry, C.I., D.J. Martin-Robichaud, and T.J. Benfey, Hormonal sex reversal of Atlantic halibut (*Hippoglossus hippoglossus L.*). Aquaculture, 2003. **219**(1-4): p. 769-781.

157. Chourrout, D. and E. Quillet, Induced gynogenesis in the rainbow trout: sex and survival of progenies production of all-triploid populations. Theoretical and Applied Genetics, 1982. **63**(3): p. 201-205.

158. Chatain, B., E. Saillant, and S. Peruzzi, Production of monosex male populations of European seabass, Dicentrarchus labrax L. by use of the synthetic androgen 17α-methyldehydrotestosterone. Aquaculture, 1999. **178**(3-4): p. 225-234.

159. Gao, Z., et al., Gonadal sex differentiation in the bluegill sunfish Lepomis macrochirus and its relation to fish size and age. Aquaculture, 2009. **294**(1-2): p. 138-146.

160. Schill, D.J., et al., Production of a YY male brook trout broodstock for potential eradication of undesired brook trout populations. North American Journal of Aquaculture, 2016. **78**(1): p. 72-83.

161. Palaiokostas, C., et al., A new SNP-based vision of the genetics of sex determination in European sea bass (*Dicentrarchus labrax*). Genetics Selection Evolution, 2015. **47**(1): p. 68.

162. Von Bargen, J., C.T. Smith, and J. Rueth, Development of a Chinook salmon sex identification SNP assay based on the growth hormone pseudogene. Journal of Fish and Wildlife Management, 2015. **6**(1): p. 213-219.

163. Larson, W.A., et al., Identification and characterization of sex-associated loci in sockeye salmon using genotyping-by-sequencing and comparison with a sex-determining assay based on the sdY gene. Journal of Heredity, 2016. **107**(6): p. 559-566.

164. Sato, M., et al., A highly flexible and repeatable genotyping method for aquaculture studies based on target amplicon sequencing using next-generation sequencing technology. Scientific Reports, 2019. **9**(1): p. 6904.

165. Graham, C.F., et al., Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADS eq). Molecular Ecology Resources, 2015. **15**(6): p. 1304-1315.

166. Gautier, M., et al., The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology, 2013. **22**(11): p. 3165-3178.

167. Onda, Y., et al., Multiplex PCR targeted amplicon sequencing (MTA-Seq): simple, flexible, and versatile SNP genotyping by highly multiplexed PCR amplicon sequencing. Frontiers in Plant Science, 2018. **9**: p. 201.

168. Dupuis, J.R., et al., HiMAP: Robust phylogenomics from highly multiplexed amplicon sequencing. Molecular Ecology Resources, 2018. **18**(5): p. 1000-1019.

169. Meek, M.H. and W.A. Larson, The future is now: amplicon sequencing and sequence capture usher in the conservation genomics era. Molecular Ecology Resources, 2019.

170. LUTZ‐CARRILLO, D.J., et al., Isolation and characterization of microsatellite loci for Florida largemouth bass, *Micropterus salmoides floridanus*, and other micropterids. Molecular Ecology Resources, 2008. **8**(1): p. 178-184.

171. Smith, A.J., et al., Body shape evolution in sunfishes: divergent paths to accelerated rates of speciation in the centrarchidae. Evolutionary Biology, 2015. **42**(3): p. 283-295.

172. Philipp, D.P., W.F. Childers, and G.S. Whitt, A biochemical genetic evaluation of the northern and Florida subspecies of largemouth bass. Transactions of the American Fisheries Society, 1983. **112**(1): p. 1-20.

173. Johnson, R. and T. Fulton, Persistence of Florida largemouth bass alleles in a northern Arkansas population of largemouth bass, *Micropterus salmoides* Lacepede. Ecology of Freshwater Fish, 1999. **8**(1): p. 35-42.

174. Interior, U.S.D.o., National survey of fishing, hunting, and wildlife-associated recreation. 2006.

175. Thompson, B.C., et al., Economic and conservation impacts of stocking wild Florida Bass into large Florida lakes. North American Journal of Fisheries Management, 2016. **36**(3): p. 452-464.

176. Buynak, G.L., et al., Stocking subadult largemouth bass to meet angler expectations at Carr Creek Lake, Kentucky. North American Journal of Fisheries Management, 1999. **19**(4): p. 1017-1027.

177. Bert, T.M., et al., Genetic management of hatchery-based stock enhancement, in Ecological and genetic implications of aquaculture activities. 2007, Springer. p. 123-174.

178. Austin, J.D., et al., An assessment of hatchery effects on Florida bass (*Micropterus salmoides floridanus*) microsatellite genetic diversity and sib-ship reconstruction. Aquaculture Research, 2012. **43**(4): p. 628-638.

179. Hargrove, J.S. and J.D. Austin, Parentage and mating patterns in a Florida Largemouth Bass (*Micropterus salmoides floridanus*) hatchery. Aquaculture Research, 2017. **48**(6): p. 3272-3277.

180. Lutz-Carrillo, D.J., et al., Admixture analysis of Florida largemouth bass and northern largemouth bass using microsatellite loci. Transactions of the American Fisheries Society, 2006. **135**(3): p. 779-791.

181. Seyoum, S., et al., Isolation and characterization of eighteen microsatellite loci for the largemouth bass, *Micropterus salmoides*, and cross amplification in congeneric species. Conservation Genetics Resources, 2013. **5**(3): p. 697-701.

182. Beck, B.H., et al., Hepatic transcriptomic and metabolic responses of hybrid striped bass (*Morone saxatilis*× *Morone chrysops*) to acute and chronic hypoxic insult. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics, 2016. **18**: p. 1-9.

183. Quist, M.C., et al., Ecology of Larval White Bass in a Large Kansas Reservoir. North American Journal of Fisheries Management, 2002. **22**(2): p. 637-642.

184. Fuller, S.A., et al., Hybrid striped bass National Breeding Program: Research towards genetic improvement of a non-model species. Bulletin of Fisheries Research Agency, 2017. **45**: p. 89-100.

185. Garber, A.F. and C.V. Sullivan, Selective breeding for the hybrid striped bass (*Morone chrysops, Rafinesque*× *M. saxatilis, Walbaum*) industry: status and perspectives. Aquaculture Research, 2006. **37**(4): p. 319-338.

186. Hallerman, E.M., Toward Coordination and Funding of Long‐Term Genetic Improvement Programs for Striped and Hybrid Bass Morone sp. Journal of the World Aquaculture Society, 1994. **25**(3): p. 360-365.

187. McEntire, M., et al., Effect of contrasting agents on survival, performance, and condition of larval hybrid striped bass *Morone chrysops* x *M. saxatilis* in tanks. Journal of Applied Aquaculture, 2015. **27**(1): p. 1-28.

188. Carlberg, J., et al., US production and sales of hybrid striped bass 1987-2004. 2004.

189. Fuller, S.A. and M.M. McEntire, Variation in body weight and total length among families of white bass, Morone chrysops, fry after communal rearing. Journal of Applied Aquaculture, 2011. **23**(3): p. 250-255.

190. Kohler, C.C., et al., Habituation to captivity and controlled spawning of white bass. Transactions of the American Fisheries Society, 1994. **123**(6): p. 964-974.

191. Kohler, C.C., et al., Performance comparison of geographic strains of white bass (Morone chrysops) to produce sunshine bass. Aquaculture, 2001. **202**(3-4): p. 351-357.

192. Fuller, S.A., et al., Heritability of growth traits and correlation with hepatic gene expression among hybrid striped bass exhibiting extremes in performance. Cogent Biology, 2018. **4**(1): p. 1453319.

193. Li, C., et al., Transcriptome annotation and marker discovery in white bass (*Morone chrysops*) and striped bass (*Morone saxatilis*). Animal Genetics, 2014. **45**(6): p. 885-887.

194. Robinson, M.C. and S.C. Courtenay, Genetic investigations on striped bass (Morone saxatilis) in the Canadian Maritime provinces. 1999: Fisheries and Oceans Canada.

195. Roy, N.K., L. Maceda, and I. Wirgin, Isolation of microsatellites in striped bass *Morone saxatilis* (Teleostei) and their preliminary use in population identification. Molecular Ecology, 2000. **9**(6): p. 827-829.

196. Leblanc, N.M., et al., Evidence of a genetically distinct population of Striped Bass within the Saint John River, New Brunswick, Canada. North American Journal of Fisheries Management, 2018. **38**(6): p. 1339-1349.

197. Wirgin, I.I., R. Proenca, and J. Grossfield, Mitochondrial DNA diversity among populations of striped bass in the southeastern United States. Canadian Journal of Zoology, 1989. **67**(4): p. 891-907.

198. Wirgin, I., et al., An evaluation of introgression of Atlantic coast striped bass mitochondrial DNA in a Gulf of Mexico population using formalin - preserved museum collections. Molecular Ecology, 1997. **6**(10): p. 907-916.

199. Wirgin, I., et al., Use of mitochondrial DNA polymorphisms to estimate the relative contributions of the Hudson River and Chesapeake Bay striped bass stocks to the mixed fishery on the Atlantic coast. Transactions of the American Fisheries Society, 1993. **122**(5): p. 669-684.

200. Stellwag, E.J., E.S. Payne, and R.A. Rulifson, Mitochondrial DNA diversity of Roanoke River striped bass. Transactions of the American Fisheries Society, 1994. **123**(3): p. 321-334.

201. Wirgin, I.I., et al., Mixed-stock analysis of Atlantic coast striped bass (*Morone saxatilis*) using nuclear DNA and mitochondrial DNA markers. Canadian Journal of Fisheries and Aquatic Sciences, 1997. **54**(12): p. 2814-2826.

202. Bielawski, J.P. and D.E. Pumo, Randomly amplified polymorphic DNA (RAPD) analysis of Atlantic Coast striped bass. Heredity, 1997. **78**(1): p. 32.

203. Waldman, J.R., R.E. Bender, and I.I. Wirgin, Multiple population bottlenecks and DNA diversity in populations of wild striped bass, Morone saxatilis. Fishery Bulletin National Oceanic and Atmospheric Adminstration, 1998. **96**: p. 614-620.

204. Diaz, M., et al., Striped bass population subdivision within the Santee-Cooper system, South Carolina. Molecular Marine Biology and Biotechnology, 1998. **7**: p. 191-196.

205. Han, K., et al., Isolation and characterization of microsatellite loci for striped bass (*Morone saxatilis*). Marine Biotechnology, 2000. **2**(5): p. 405-408.

206. Brown, K., et al., Isolation and characterization of nuclear microsatellite loci in the anadromous marine fish Morone saxatilis. Molecular Ecology Notes, 2003. **3**(3): p. 414-416.

207. Brown, K., G. Baltazar, and M. Hamilton, Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay striped bass (*Morone saxatilis*). Heredity, 2005. **94**(6): p. 606.

208. Chapman, R.W., Mitochondrial DNA analysis of striped bass populations in Chesapeake Bay. Copeia, 1990: p. 355-366.

209. Laughlin, T.F. and B.J. Turner, Hypervariable DNA markers reveal high genetic variability within striped bass populations of the lower Chesapeake Bay. Transactions of the American Fisheries Society, 1996. **125**(1): p. 49-55.

210. Diaz, M., G.M. Leclerc, and B.E. Fishtec, Nuclear DNA markers reveal low levels of genetic divergence among Atlantic and Gulf of Mexico populations of striped bass. Transactions of the American Fisheries Society, 1997. **126**(1): p. 163-165.

211. White, M.M., Genetic variation in white bass. Transactions of the American Fisheries Society, 2000. **129**(3): p. 879-885.

212. Couch, C., et al., Isolation and characterization of 149 novel microsatellite DNA markers for striped bass, *Morone saxatilis*, and cross‐species amplification in white bass, Morone chrysops, and their hybrid. Molecular Ecology Notes, 2006. **6**(3): p. 667-669.

213. Haponski, A.E. and C.A. Stepien, Phylogenetic and biogeographical relationships of the Sander pikeperches (Percidae: Perciformes): patterns across North America and Eurasia. Biological Journal of the Linnean Society, 2013. **110**(1): p. 156-179.

214. Bozek, M.A., D.A. Baccante, and N.P. Lester, Walleye and sauger life history. Biology, management, and culture of Walleye and Sauger, 2011. **233**: p. 301.

215. Christie, G.C. and H.A. Regier, Measures of optimal thermal habitat and their relationship to yields for four commercial fish species. Canadian Journal of Fisheries and Aquatic Sciences, 1988. **45**(2): p. 301-314.

216. Zhao, Y., B.J. Shuter, and D.A. Jackson, Life history variation parallels phylogeographical patterns in North American walleye (*Sander vitreus*) populations. Canadian Journal of Fisheries and Aquatic Sciences, 2008. **65**(2): p. 198-211.

217. Schmalz, P.J., et al., Harvest and exploitation. Biology, management, and culture of walleye and sauger. American Fisheries Society, Bethesda, Maryland, 2011: p. 375-401.

218. McMahon, T.E. and D.H. Bennett, Walleye and northern pike: boost or bane to northwest fisheries? Fisheries, 1996. **21**(8): p. 6-13.

219. Jennings, M.J., et al., Evaluation of supplemental walleye stocking in northern Wisconsin lakes. North American Journal of Fisheries Management, 2005. **25**(3): p. 1171-1178.

220. Seddon, P.J., D.P. Armstrong, and R.F. Maloney, Developing the science of reintroduction biology. Conservation Biology, 2007. **21**(2): p. 303-312.

221. Waterhouse, M.D., B.L. Sloss, and D.A. Isermann, Relationships among Walleye population characteristics and genetic diversity in northern Wisconsin lakes. Transactions of the American Fisheries Society, 2014. **143**(3): p. 744-756.

222. Allen, B.E., et al., Loss of SNP genetic diversity following population collapse in a recreational walleye (*Sander vitreus*) fishery. Canadian Journal of Fisheries and Aquatic Sciences, 2017. **75**(10): p. 1644-1651.

223. Haponski, A.E. and C.A. Stepien, Two decades of genetic consistency in a reproductive population in the face of exploitation: patterns of adult and larval walleye (*Sander vitreus*) from Lake Erie’s Maumee River. Conservation genetics, 2016. **17**(6): p. 1345-1362.

224. Garner, S.R., S.M. Bobrowicz, and C.C. Wilson, Genetic and ecological assessment of population rehabilitation: Walleye in Lake Superior. Ecological applications, 2013. **23**(3): p. 594-605.

225. Stepien, C.A., O.J. Sepulveda-Villet, and A.E. Haponski, Comparative genetic diversity, population structure, and adaptations of walleye and yellow perch across North America, in Biology and Culture of Percid Fishes. 2015, Springer. p. 643-689.

226. Coykendall, D.K., et al., Development of eighteen microsatellite loci in walleye (*Sander vitreus*). Conservation Genetics Resources, 2014. **6**(4): p. 1019-1021.

227. Haponski, A.E. and C.A. Stepien, A population genetic window into the past and future of the walleye Sander vitreus: relation to historic walleye and the extinct “blue pike” S. v.“glaucus”. BMC Evolutionary Biology, 2014. **14**(1): p. 133.

228. Strange, R.M. and C.A. Stepien, Genetic divergence and connectivity among river and reef spawning groups of walleye (*Sander vitreus vitreus*) in Lake Erie. Canadian Journal of Fisheries and Aquatic Sciences, 2007. **64**(3): p. 437-448.

229. Stepien, C.A., et al. Status and delineation of walleye (*Sander vitreus*) genetic stock structure across the Great Lakes. in Status of walleye in the Great Lakes: proceedings of the 2006 symposium. Great Lakes Fishery Commission technical report. 2010.

230. Stepien, C.A., et al., Signatures of vicariance, postglacial dispersal and spawning philopatry: population genetics of the walleye *Sander vitreus*. Molecular Ecology, 2009. **18**(16): p. 3411-3428.

231. Stepien, C.A., et al., Temporal and spatial genetic consistency of walleye spawning groups. Transactions of the American Fisheries Society, 2012. **141**(3): p. 660-672.

232. Stepien, C.A. and J.E. Faber, Population genetic structure, phylogeography and spawning philopatry in walleye (*Stizostedion vitreum*) from mitochondrial DNA control region sequences. Molecular Ecology, 1998. **7**(12): p. 1757-1769.

233. Haponski, A.E. and C.A. Stepien, Genetic connectivity and diversity of walleye (*Sander vitreus*) spawning groups in the Huron–Erie Corridor. Journal of Great Lakes Research, 2014. **40**: p. 89-100.

234. Billington, N., et al., Phylogenetic relationships among four members of Stizostedion (Percidae) determined by mitochondrial DNA and allozyme analyses. Journal of Fish Biology, 1991. **39**: p. 251-258.

235. Billington, N., R.J. Barrette, and P.D. Hebert, Management implications of mitochondrial DNA variation in walleye stocks. North American Journal of Fisheries Management, 1992. **12**(2): p. 276-284.

236.Billington, N. Geographical distribution of mitochondrial DNA (mtDNA) variation in walleye, sauger, and yellow perch. in Annales Zoologici Fennici. 1996. JSTOR.

237. Billington, N., P.D. Hebert, and R.D. Ward, Allozyme and mitochondrial DNA variation among three species of Stizostedion (Percidae): phylogenetic and zoogeographical implications. Canadian Journal of Fisheries and Aquatic Sciences, 1990. **47**(6): p. 1093-1102.

238. Bickford, N. and R. Hannigan, Stock identification of walleye via otolith chemistry in the Eleven Point River, Arkansas. North American Journal of Fisheries Management, 2005. **25**(4): p. 1542-1549.

239. White, M.M., et al., A genetic assessment of Ohio River walleyes. Transactions of the American Fisheries Society, 2005. **134**(3): p. 661-675.

240. Jennings, M.J. and D.P. Philipp, Use of allozyme markers to evaluate walleye stocking success. North American Journal of Fisheries Management, 1992. **12**(2): p. 285-290.

241. Ward, R.D., N. Billington, and P.D. Hebert, Comparison of allozyme and mitochondrial DNA variation in populations of walleye, *Stizostedion vitreum.* Canadian Journal of Fisheries and Aquatic Sciences, 1989. **46**(12): p. 2074-2084.

242.Colby, P.J. and S.J. Nepszy, Variation among stocks of walleye (*Stizostedion vitreum vitreum*): management implications. Canadian Journal of Fisheries and Aquatic Sciences, 1981. **38**(12): p. 1814-1831.

243. Hedrick, P.W., Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution, 1999. **53**(2): p. 313-318.

244. Billington, N. and R.M. Strange, Mitochondrial DNA analysis confirms the existence of a genetically divergent walleye population in northeastern Mississippi. Transactions of the American Fisheries Society, 1995. **124**(5): p. 770-776.

245. Billington, N., R.M. Strange, and M.J. Maceina. Mitochondrial-DNA confirmation of southern walleye in the Mobile Basin, Alabama. in Proceedings of the Annual Conference Southeastern Association of Fish and Wildlife Agencies. 1997.

246. Chen, K.-Y., Lake Erie walleye population structure and stock discrimination methods. 2016, The Ohio State University.

Chapter II SNP marker panels for parentage assignment and traceability in the Florida bass (*Micropterus floridanus*)

Abstract

The Florida bass (*Micropterus floridanus*) is a species endemic to peninsular Florida that is held in high esteem by bass anglers for its tendency to attain a larger maximum size and aggressiveness relative to that of its sister taxon, the Northern largemouth bass, *Micropterus salmoides*. Hatchery rearing and stocking of Florida bass outside of their native range are commonplace, particularly in the southern United States. In many cases, however, there has been minimal assessment of the persistence and success of these fish. Genetic markers are an important tool for tagging and tracing the contributions of particular lines and crosses of fish. Single nucleotide polymorphism (SNP) markers, in particular, can provide rapid and affordable genotyping of large numbers of fish. In the present study, I generated 58,450 genome-wide SNPs and population-level genotypes for Florida bass using a cost-effective genotyping-by-sequencing method. A total of 58 SNPs were shown to assign parents to offspring with 100% accuracy, irrespective of sex and with the presence of full-sib relationships. Depending on the population, sex information, and genetic relationships between parents, I also demonstrated that smaller SNP subsets may be sufficient for parentage assignment. The accuracy and assignment power of the SNP panels were found to compare favorably to those of 10 microsatellites genotyped on the same parents and progeny. This study demonstrated the utility of simple and low-cost GBS techniques for SNP discovery and the relatively small number of variable SNPs needed for accurate parentage assignment in Florida bass. The SNP resources created in this study should facilitate parentage-based research and breeding, genetic tagging, and conservation of Florida bass.

1. Introduction

The artificial propagation of aquatic species and subsequent release into natural environments, also known as stock enhancement, has been a widely utilized and frequently criticized approach in conservation and supplementation efforts [1-5]. Additional forms of stocking involve introducing non-native species (e.g. Florida bass *Micropterus floridanus*) to enhance specific fisheries attributes (e.g. growth; [6, 7]). One of the challenges in stock enhancement is to maintain pedigree information for hatchery brood individuals. Reliable pedigree information allows fisheries managers to track ecological or life-history characteristics of released fish, to estimate genetic parameters and breeding values, and to minimize inbreeding in broodstocks [2, 8]. The predominant approaches for pedigree development and hatchery stocks evaluation in aquaculture involve the use of physical tags to determine the origin and age of recaptured fish [9, 10]. However, there are known drawbacks to these traditional tagging techniques including tissue damage, decreased swimming capacity, premature tag loss, and risk related to juvenile handling vulnerability [11]. Therefore, alternative tagging techniques are needed. One emerging technology is the use of parentage-based tagging (PBT), a genetic-based tagging method, to create a database of parental genotypes from hatcheries and later assign each progeny back to their parents, thereby reconstructing the pedigree and identifying the origin and brood year for each sampled offspring [10]. The implementation of large-scale PBT project in steelhead (*Oncorhynchus mykiss*) has demonstrated the feasibility of this method in building parent-offspring relationships with a number of advantages such as low cost and higher tagging rates compared with traditional tagging methods [10].

Advances in molecular technologies have allowed scientists to develop DNA markers that are polymorphic and robust for parentage analysis. In aquaculture, parentage assignment studies came of age in the 1990s with the advent of microsatellite markers [12-14]. However, this parentage assignment approach has frequently encountered issues associated with genotyping error, null alleles, and mutations that limit its resolving power [15, 16]. Single nucleotide polymorphisms (SNPs) are codominant, biallelic molecular markers valued for their genome-wide distribution, abundance, ease of multiplexing and low genotyping error rate for high-throughput analyses [17, 18]. SNPs are rapidly replacing microsatellites in parentage studies as the development of SNPs is more efficient and less expensive for nonmodel aquatic species [8, 10, 19-21]. Additionally, with advances in SNP genotyping approaches, SNPs are expected to become one of the major marker systems for routine parentage analysis in a variety of aquatic species [22]. Given that reference genomes are currently available for only a limited number of taxa, genotyping-by-sequencing (GBS) data is proposed as one of the best options for cost-effective SNP discovery and subsequent parentage studies [23]. GBS is a simple, reproducible, highly multiplexed approach that was originally developed for SNP identification and genotyping in crop genomes and populations [24]. GBS has been increasingly used for genetic and genomic research in nonmodel organisms, such as linkage map construction [25, 26], marker-assisted selection (MAS) [27], trait mapping [28], and estimating genetic diversity [29].

The Florida bass (*Micropterus floridanus*) is a highly prized sportfish native to peninsular Florida that attains larger maximum sizes relative to its sister taxon, the Northern largemouth bass, *Micropterus salmoides* [30]. The distribution of Florida bass overlaps with that of the Northern largemouth bass forming a natural hybrid zone in the southeastern US; however, the scope of this introgression has been expanded dramatically via stocking efforts [30-33]. Although Florida bass have been extensively studied from a stocking and management perspective, genetic efforts have mainly focused on the development of markers capable of assessing population structure and/or

Florida/Northern ancestry [32-35] and extralimital introduction [36]. Relatively little attention has been paid to developing tagging methods for pedigree development in this species. Recently, microsatellites have been used for parentage assignment in *M. floridanus* to investigate the mating patterns in hatchery environments [37, 38]. However, the assignment power and the minimum number of microsatellite markers needed to accurately conduct parentage analysis in Florida bass groups were not estimated. Thus, the goal of the present study was to develop SNP markers suitable for Florida bass parentage analysis through GBS followed by validation and panel creation using Agena MassARRAY technology. Further, I evaluated the accuracy of SNP markers in parentage analysis relative to previously utilized microsatellite markers. And lastly, I compared the performance of multiple parentage programs (Cervus and SNPPIT; [16, 39]). The comprehensive assessment should provide valuable information for investigators developing SNPs via GBS for nonmodel organisms, and the SNP resources reported here should be of high utility in pedigree tracing of hatchery-reared Florida bass used for stock enhancement or a genetic selection program.

2. Methods and Materials

2.1 Genotyping-by-Sequencing Sample Preparation and Sequencing

A total of 265 hatchery-reared Florida bass were collected for library construction and downstream GBS analysis. Among these fish, three families of 250 individuals were collected at the Go Fish Education Center (GFEC), Perry GA and the other 15 Florida bass samples were collected from St. Johns River in Florida (STJR). Samples were confirmed genetically as Florida bass using a previously published panel of 25 diagnostic SNP makers [32], as well as 38 additional SNP markers validated subsequently using the same methods and reference samples as Li et al. (2015) [32].

Genomic DNA from all samples was extracted from blood or fin clips using the DNeasy Blood & Tissue kit (Qiagen, Valencia, CA) according to the manufacturer's protocol. DNA quality was assessed by running 100 ng of each DNA sample on 1% agarose gels. DNA concentration was determined using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen). DNA samples were sent to the Institute for Genomic Diversity, Cornell University, for sequencing. Library preparation and sequencing protocols closely followed those described in Elshire et al. (2011) [24], with minor modifications. Briefly, oligonucleotides comprising the top and bottom strands of each barcode adapter and a common adapter were diluted (separately) in TE (50 mM each) and annealed in a thermocycler. Barcode and common adapters were quantified and diluted in water, mixed together in a 1:1 ratio and aliquoted into a 96-well PCR plate and dried down. DNA samples (100 ng in a volume of 10 μL) were added to individual adapter-containing wells and plates were dried. Samples (DNA plus adapters) were digested with *Pst*I (CTGCAG) using the manufacturer's recommended conditions. *Pst*I is a 6-bp cutting enzyme that targets and cuts regions of the genome containing little repetitive DNA [40]. Following digestion, adapters were then ligated to sticky ends by adding T4 ligase (New England Biolabs). Samples were incubated at 22 °C for 1 h and heated to 65 °C for 30 min to inactivate the T4 ligase. DNA samples, each with a different barcode adapter, were combined into sets of 96 samples purified using a commercial kit (QIAquick PCR Purification Kit; Qiagen, Valencia, CA), and then restriction fragments with ligated adapters were amplified to generate sequencing libraries. Libraries were purified as above and fragment sizes were evaluated on an Experion™ automated electrophoresis station (Bio-Rad, Hercules, CA). Single-end sequencing of one 96-plex library per lane was performed on an Illumina HiSeq instrument with 100-bp read chemistry [41].

2.2 DNA Sequence Analysis and SNP Marker Development

The GBS analysis pipeline (TASSEL Version 4.0; [42]) was used for tag alignment and subsequent SNP calling (http://www.maizegenetics.net/tassel). To generate reference-based SNPs and design primers for the SNP panels, I assembled a rough draft genome for largemouth bass. A DNA sample from a female Florida bass was sequenced using 150 base-length paired-end (PE) read chemistry on the Illumina HiSeq 2500. The Maryland Super Read Celera Assembler (MaSuRCA; [43]) was used for the *de novo* draft genome assembly. I followed default parameter settings with library insert average length and standard deviation set as 400 and 60, respectively. The assembled draft genome consisted of 1,001,521,225 bp distributed across 249,768 contigs and had a GC content of 41.1%. The longest contig had a length of 638,759 bp, and the N50 contig length was 11,136 bp. Using the estimate_genome_size.pl script developed by Ryan (2013) [44], it was estimated that 74% of the genome was covered by the draft assembly. During GBS analysis, sequences were first trimmed of ambiguous nucleotides and barcodes. Potential chimeric sequences were eliminated by trimming the sequence at the corresponding restriction enzyme site, if present. After removing low-quality reads, the remaining high-quality reads from each individual were mapped to the preliminary assembly of largemouth bass contigs using Burrows-Wheeler Aligner (BWA; [45]), and the nonspecific matches were discarded. Identical, aligned reads were clustered into tags. SNP discovery was performed for each set of tags that aligned to the same starting genomic position. After multiple sequence alignment, the allele represented by each tag was determined to tally the observed depths of each allele. The genotype of the SNP was then determined by a binomial likelihood ratio method of quantitative SNP calling. During pre-filtered SNP detection, the minimum minor allele frequency (MAF) was set to 0.05 (overall), the minimum minor allele count (MAC) was set to 10 (overall), and minimum locus coverage (LCov) was set to 0.1 (overall). The LCov of 0.1 ensured that at least 10% of individuals were genotyped for a SNP calling.

To ensure that the SNPs were polymorphic in diverse Florida bass populations and informative for parentage assignment, VCFtools [46] and BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) were used for stringent filtration of SNPs based on the following criteria: 1) Only biallelic SNPs were kept (-min-alleles 2 -max-alleles 2); 2) the SNPs were called in at least 90% of individuals (-max-missing 0.9); 3) SNPs with a MAF > 0.4 were preferentially selected (-maf 0.4); 4) I eliminated loci with quality score < 98 (-minGQ 98), a stringent filtering criteria considering genotype confidence and sequence depth per locus [47]; 5) SNPs not deviating from Hardy-Weinberg equilibrium (HWE) with p-value > 0.05 were kept (STJR was not included due to limited sample size; -hwe 0.05); 6) To improve the quality of subsequent assay design, I required that there were no adjacent SNPs within 30 bp on each side of the targeted SNP because tightly linked SNPs may potentially alter primer binding; 7) Only SNPs with at least a 100-bp flanking region on either side of the polymorphic site were selected for the assay design (required by MassARRAY Assay Design Software); 8) Due to the lack of genetic maps at the time during assay design, I only selected SNPs with flanking regions (201 bp) which could be mapped to unique location without mismatched bases in the draft genome based on BLAST hit (e-value ≤ 1E-5).

2.3 SNP and Microsatellite Genotyping

A MassARRAY System (Agena Bioscience, San Diego, CA) was employed to validate a subset of SNPs identified and genotyped by GBS and to evaluate the performance of selected SNPs across diverse bass populations. Samples utilized for the GBS sequencing, as well as new samples from additional populations, were used for genotyping. These include 16 fish individuals from DKLA (Dekalb County Lake, AL), 264 from natural waterbodies near the FBCC (Florida Bass Conservation Center, Webster, FL), 100 from ALLA (Lake Allatoona, GA), and 16 from OLHR (Old Hickory Reservoir, TN). SNP assays were designed using the MassARRAY Assay Design Software with the goal of maximizing multiplexing of 40 SNPs per well. Using MassARRAY Assay Design Software, I designed and ordered two multiplex assays, with 40 and 31 SNPs, respectively (Table 1). Amplification and extension reactions were performed using 10 ng of DNA per sample and utilizing the iPLEX Gold Reagent Kit according to the manufacturer's protocols. SNP genotypes were called using the MassARRAY Typer 4 analysis software. This software uses a three-parameter (mass, peak height and signal-to-noise ratio) model to calculate the significance of each genotype. A final genotype was called and assigned a particular name (e.g. conservative, moderate, aggressive, user call) based on probability. Noncalls were also noted (e.g. low probability, bad spectrum).

Table 1 Primer sequence information for the two MassARRAY multiplex panels in Florida bass.

SNP_ID	SNP		Primer Sequences
S102090_10541	C/A	PCR1:	ACGTTGGATGGCAATACAGTTCAGCGAAAC
		PCR2:	ACGTTGGATGACTGCAGATCTACAGTCCAG
		EXT:	AGAGATGTGACTTTCTTCATTCA
S102873_2971	T/A	PCR1:	ACGTTGGATGCACGCAGGACCTTAACTTTG
		PCR2:	ACGTTGGATGCTGCACTAAAACTCCTTCCC
		EXT:	ACCTTAACTTTGACCACC
S103591_2492	A/T	PCR1:	ACGTTGGATGGAACAACAGCTGCAGCTAAC
		PCR2:	ACGTTGGATGAGTGGAAGCCACTTCAGAAC
		EXT:	AGAACATGAGTGAATCAGTA
S106802_2168	A/G	PCR1:	ACGTTGGATGAGGGTTCTGCTGCAGATAAG
		PCR2:	ACGTTGGATGTAACACACTGCAGCATGAAC
		EXT:	GCAGCATGAACAAGACATT
S116433_8507	G/T	PCR1:	ACGTTGGATGGCGAGGACTTAACAGAAAGC
		PCR2:	ACGTTGGATGGCATTCATGTAGAGCTTTCC

SNP_ID	SNP		Primer Sequences
		EXT:	cGCTTTCCCTCTAACCA
S124943_1970*	G/A	PCR1:	ACGTTGGATGGCAGTCCTTGCATTTGTCAG
		PCR2:	ACGTTGGATGGCATTTGTCTTTCTGCAGGG
		EXT:	cctccTCTGCAGCTTTCCTAAC
S137063_6774	G/C	PCR1:	ACGTTGGATGATTGTCATACGTCTGCCAGC
		PCR2:	ACGTTGGATGCTGCAGATATGAGATGTTGG
		EXT:	cttTATCACAACACACGCAC
S147389_2312	T/C	PCR1:	ACGTTGGATGTGCTTTCTTTCCCCCACTTC
		PCR2:	ACGTTGGATGTGGCTGTGTTTTAGCTTTGG
		EXT:	gtgaTGCAGTTAAGCGTCTCTGGC
S152803_15570	G/A	PCR1:	ACGTTGGATGGCACATCTGCAGCCTGTTTT
		PCR2:	ACGTTGGATGATGGCTGCGTGCCATTCTAC
		EXT:	aTGCAGCCTGTTTTTATCCTG
S180270_68030	C/T	PCR1:	ACGTTGGATGGAGATGGATCTGCAGTGATG
		PCR2:	ACGTTGGATGCCCTCCATTCAAGCTGATAC
		EXT:	TTTCCTGCCTGTTCC
S182265_3231	G/A	PCR1:	ACGTTGGATGTTTCAGACAGTGTCTGGAGC
		PCR2:	ACGTTGGATGCTTCAGGTGCTGCAGCTCT
		EXT:	tgcccGCAGGGTCTCCTCTCCCTC
S182533_8441	T/G	PCR1:	ACGTTGGATGCATTGATGCACACACATGCG
		PCR2:	ACGTTGGATGACAGACTTGAGCTGCACCTG
		EXT:	ctcttCAGGTTTCCCCTGACTC
S184537_4064	T/C	PCR1:	ACGTTGGATGAGGTACGGAGCGGCATTTAG
		PCR2:	ACGTTGGATGACAGGAGGACGGATCAATCG
		EXT:	ggtaGGCATTTAGTCCCGCAGAGC
S184629_511	G/A	PCR1:	ACGTTGGATGTATCATTTCTGAGCTGCAGG
		PCR2:	ACGTTGGATGATGTGCAGCTGCAGTGAATC
		EXT:	ttatCTGCAGGTTATAAACACTC
S189438_3105	C/T	PCR1:	ACGTTGGATGAAGTCAGCCACAGTCATCTC
		PCR2:	ACGTTGGATGTTTCCTTTCCTGCAGGTGTC
		EXT:	aaagCGCACCTATTGACCATT
S193945_13226	G/A	PCR1:	ACGTTGGATGAATATGTGGCTGCAGCAGAG
		PCR2:	ACGTTGGATGTATGCTTCACTGTTGCCTGC
		EXT:	CTGGAGAGCGATTCAAG
S197859_8054	A/C	PCR1:	ACGTTGGATGGCAGAAGCTGCAAATTCTGG
		PCR2:	ACGTTGGATGATTTGTAGGCCAGGGCATCC
		EXT:	CAGTACTTTATTATAATGCTTTTCT
S198419_11760*	G/A	PCR1:	ACGTTGGATGGTGTCTAAGCCCAGTGGATG
		PCR2:	ACGTTGGATGACGGCTCGTGCAGTGTTTTT
		EXT:	TTTATTATTCAGTGTAATTTTAAACTC
S200775_2909*	T/C	PCR1:	ACGTTGGATGTGTCCTGTGTGAGCTGATTC
		PCR2:	ACGTTGGATGGTATTTCCTGCAGCGTAGTG
		EXT:	ggggtCTGATTCTGCAGAATTATGA

SNP_ID	SNP		Primer Sequences
S206110_2535	C/T	PCR1:	ACGTTGGATGTCCTGACAGCTGATGAGAAC
		PCR2:	ACGTTGGATGGAGATTATCTGCTCTGGCTG
		EXT:	tggaTGCAGTCTGAGCTTC
S207601_1046	T/G	PCR1:	ACGTTGGATGTGGTGTTAGTATCTGCAGCG
		PCR2:	ACGTTGGATGGGCTTGTCACATCATCACAG
		EXT:	ggagCTGCAGCGTAATGATACT
S217958_6263	T/C	PCR1:	ACGTTGGATGGCTTCTGGTTTTTCCTACGG
		PCR2:	ACGTTGGATGAGCAACTCATCGCACAACTG
		EXT:	CACGGAGCTTCCCAC
S218919_7782	T/C	PCR1:	ACGTTGGATGTCATGACTCGGAGGACGTTC
		PCR2:	ACGTTGGATGCTCCTCTCTCTCAGCTGTTG
		EXT:	aCAGCTGTTGGACGAC
S221442_25960	A/G	PCR1:	ACGTTGGATGTTTAATCCCCTGCTGCAGAG
		PCR2:	ACGTTGGATGATTGGACTGCAGTCTGAACG
		EXT:	caAACTGGCATGGGATG
S223782_10510	A/T	PCR1:	ACGTTGGATGCAGGTACTGCAGGGCGCTGGA
		PCR2:	ACGTTGGATGAGTTTCTGCAGCACCAGCC
		EXT:	tCCTCGCCCTGCAGCCTC
S228984_5277	A/C	PCR1:	ACGTTGGATGAAAGCGTCACACTGAGTCAC
		PCR2:	ACGTTGGATGAAGGCAGATAAAGCTTCTCC
		EXT:	cggGACACCTGATGCCTGTTG
S240568_11866	G/T	PCR1:	ACGTTGGATGGTATGCACACACACACTAGC
		PCR2:	ACGTTGGATGAGATAGGAGGAAACCTCTGC
		EXT:	ccTGTGCTTGCCTCTGT
S241381_18427	A/G	PCR1:	ACGTTGGATGGTTCAGGCCCCAGTCATTTC
		PCR2:	ACGTTGGATGCGATGAGTTTATTCCTCTGG
		EXT:	TCCTCTGGGAGCTGTG
S247388_9639*	T/C	PCR1:	ACGTTGGATGATTTTCTGCAGAAGCCGTCC
		PCR2:	ACGTTGGATGGACATCCACTGCAACTCCTG
		EXT:	TCGTAGACTGCTGGA
S248643_4645*	C/T	PCR1:	ACGTTGGATGGAGAATCACAAGGACAGACC
		PCR2:	ACGTTGGATGGTTCTCTATGACACTGCAGG
		EXT:	ttcccACTGCAGATTCGCTGATC
S99816_4132	T/A	PCR1:	ACGTTGGATGACTGTAGCTCCAGTGTTTGC
		PCR2:	ACGTTGGATGCTGCAGGAGGCTGTTTAATC
		EXT:	GAGTTTCAGTCTGCACCA
S101377_1427	G/A	PCR1:	ACGTTGGATGACAGAGTGTGGAAGTATCCC
		PCR2:	ACGTTGGATGACTGGAAAAACTTGACCCCG
		EXT:	CCGTTACCTCATCGC
S101976_10904	G/C	PCR1:	ACGTTGGATGTTTGGCACATTGTTGGTAAC
		PCR2:	ACGTTGGATGCTGCTGCTGCTGTTATTCAC
		EXT:	taaaTTATCTGCAGGCTGAGACAAAC
S102040_4040	A/G	PCR1:	ACGTTGGATGGGGTCTCTGTGTAGTTCTC
		PCR2:	ACGTTGGATGTGCTAGATGCCAAAGCCAAC

SNP_ID	**SNP**		**Primer Sequences**
		EXT:	cccGTGTAGTTCTCCATCCC
S113175_2051	G/A	PCR1:	ACGTTGGATGGCAGTGTGAGGTGAGAAATC
		PCR2:	ACGTTGGATGGCAGCAGGACACATGTTGAA
		EXT:	ccgtgGAGGTGAGAAATCCAAGAAAAAT
S120573_1380*	G/C	PCR1:	ACGTTGGATGTGTTTTTCACCCTGCAGCAC
		PCR2:	ACGTTGGATGCTGCAGACAGAGATCATGTG
		EXT:	gcGCCCACACTCACACA
S124795_2087	A/G	PCR1:	ACGTTGGATGTTACGTTCTGCAGCATCCAC
		PCR2:	ACGTTGGATGGGCTGTTGAAGTTGTTCTGC
		EXT:	TGGGCATTACTATTGATGACAAGC
S126695_427	C/G	PCR1:	ACGTTGGATGATCCTGAATTTGTTGCTTCG
		PCR2:	ACGTTGGATGCAATGGCCAGGTCTATGAAG
		EXT:	gtgGCTTCGAAAGTAGCAATGA
S180294_27034	A/G	PCR1:	ACGTTGGATGTTTTCCATCGCAGCTGGTTG
		PCR2:	ACGTTGGATGGCTGCAGGGTCAAAATTGTC
		EXT:	TGTCCAGCCAACATC
S180381_7948	T/C	PCR1:	ACGTTGGATGCTCGTTGACTCATTCACTGC
		PCR2:	ACGTTGGATGTTAGGTGCTGCAGAGCAATC
		EXT:	tgttCTGCAGAGCAATCGTCTCG
S180545_29172	T/A	PCR1:	ACGTTGGATGTACAGAACAGTGACTGCAGG
		PCR2:	ACGTTGGATGGCCAAAGAGTTCTCTCCATC
		EXT:	aacgtCAGGGGAAGAAGAGCGCGTG
S180613_6006	A/G	PCR1:	ACGTTGGATGTGCTTGACTAACTGCTCCAC
		PCR2:	ACGTTGGATGAAAAGGCCAGCCTGATGGAC
		EXT:	cctgTATATATAAAACAAAAGCCACAAC
S181461_11655*	T/C	PCR1:	ACGTTGGATGTCAGAGATCAGCTGTGCTTC
		PCR2:	ACGTTGGATGCCTGCAGCTTATTGCTACTC
		EXT:	gggaGCCTCAGCTCTGAAGT
S183217_15036	A/C	PCR1:	ACGTTGGATGCACTGCAGGATAATCGTCAG
		PCR2:	ACGTTGGATGACCTGAGGAAAACTGACAGC
		EXT:	acgaAGGATAATCGTCAGGATAAC
S183224_19713*	T/C	PCR1:	ACGTTGGATGGTGTCAGATCCTCTTTGTGC
		PCR2:	ACGTTGGATGTGTAACTGAAACCTGCAGCG
		EXT:	GCTGCAGGACAAACACAACT
S185066_8773	A/C	PCR1:	ACGTTGGATGCTGATGAACCTGACTGTCTC
		PCR2:	ACGTTGGATGGAAACCTGCTGCAGTCAAAG
		EXT:	acgacACCTGCTGCAGTCAAAGTGGATC
S185813_2788	G/A	PCR1:	ACGTTGGATGTGCAGACATGATTGGGCTGG
		PCR2:	ACGTTGGATGACTGCAGCATGGGTAAAGGC
		EXT:	ctcccCTGCTGCACTGTGCAAC
S186135_8412	T/A	PCR1:	ACGTTGGATGTGCTCACTGTAGCAGATGTC
		PCR2:	ACGTTGGATGTGTCATCTGGTCGTGGTTTG
		EXT:	cgcTCTGAATCCTGCAGTC
S187068_5343	T/C	PCR1:	ACGTTGGATGCCTGTAGAGCAATTCAGACC

SNP_ID	SNP		Primer Sequences
		PCR2:	ACGTTGGATGCTGCAGCGGGAAAAAATAAC
		EXT:	ggacATGTTAATCTGCGGCTGTCAGTC
S189228_26225	G/A	PCR1:	ACGTTGGATGACATCTGCCTTAATAGCTTG
		PCR2:	ACGTTGGATGCAGCTGGTAAGATCTGATAT
		EXT:	ggaatGGATGAGGTCCATAATGA
S189846_454	C/T	PCR1:	ACGTTGGATGTGTTCCCTGCACTCAATCAC
		PCR2:	ACGTTGGATGCCATCATGGTAAGATGGCTG
		EXT:	tctaCCCAGAGTGACTCTTCGCGTG
S196576_33262	A/C	PCR1:	ACGTTGGATGAGTCACACACCTGCAGTATG
		PCR2:	ACGTTGGATGTGACCGAGCTGATGAACAAC
		EXT:	gggtGTCATCGGGTACGCCGTGGTC
S209202_19859	C/T	PCR1:	ACGTTGGATGTCATCTGCAGCACAACTGTC
		PCR2:	ACGTTGGATGGCCTGCAGCTTTCCTTAAAC
		EXT:	acaaAACTGTCTTGTCAATATTATCAAC
S212194_31242	T/A	PCR1:	ACGTTGGATGGGTCTAACTGAATCACACGC
		PCR2:	ACGTTGGATGAACAGCAGGTCGATGAACAG
		EXT:	AGCATCTGCAGATAAGC
S218414_5842*	G/A	PCR1:	ACGTTGGATGGAGATGGATCTGCAGTGATG
		PCR2:	ACGTTGGATGACCTGCCTTCCATACAGAAC
		EXT:	CCCGTTTACTGACTCTG
S220253_1360	A/C	PCR1:	ACGTTGGATGTTCGCCTCAAATTGTCCTGC
		PCR2:	ACGTTGGATGTGCAGGAGCTAAAGTGAAGG
		EXT:	ccccTGGCTGCAGCTTCTCCAA
S220499_32696	G/T	PCR1:	ACGTTGGATGCTGCAGGTTTAGTGAAGAGG
		PCR2:	ACGTTGGATGCTAACCATCTGCAGGCTGGA
		EXT:	gagtCAGGCTGGAGGTATACT
S224755_5094*	G/C	PCR1:	ACGTTGGATGGTCACTCAACATGCATGTGG
		PCR2:	ACGTTGGATGTAAGTGTACATTTCCCCTGC
		EXT:	ccTTTCCCCTGCAGTAATGAACAAGTT
S227350_17628	C/G	PCR1:	ACGTTGGATGCAGGGAGGCAAAACATACAC
		PCR2:	ACGTTGGATGCTCCAGCGACAAGGTAAGAG
		EXT:	ccctTGCAGGGCCCCAGGCCAAAC
S229452_24231	A/G	PCR1:	ACGTTGGATGGGTAAACCCAAACTGCAGAC
		PCR2:	ACGTTGGATGTATGAACCCAGTCTGCAGAG
		EXT:	cctgGTGGTCCCTCTCTGGTTT
S231643_22906	A/C	PCR1:	ACGTTGGATGTGCCACTGCAGGTGTTTTTG
		PCR2:	ACGTTGGATGTGCATTTGTTGTGCCTCCTC
		EXT:	TTCTGCAGCCAAGTG
S236143_10924	G/C	PCR1:	ACGTTGGATGAGGGTTTAGTTTCCCCTCAC
		PCR2:	ACGTTGGATGGCTGCAGAAACTGAAGTGAG
		EXT:	ggAGTGAGTTCAGAAAATAAAACGGA
S236993_6123	A/G	PCR1:	ACGTTGGATGGATCAACAGGACCAGACACC
		PCR2:	ACGTTGGATGACCAGCAGAGTGAGTCAAAG
		EXT:	ggGATTCTGCTGCAGTGA

SNP_ID	SNP		Primer Sequences
S237520_3355*	A/G	PCR1:	ACGTTGGATGCAGCTGACTGCAGAAAAGAC
		PCR2:	ACGTTGGATGGACTCTGCAGAGTACTAATG
		EXT:	aTTGAAAGCTGCAGGAAT
S237730_11307	T/A	PCR1:	ACGTTGGATGTGCGTGTCATCCTGAAACTC
		PCR2:	ACGTTGGATGCATTAATGAGGTTCCTCTCG
		EXT:	acccgTTTTAAGGTGTGTCTGAGTGCG
S240894_43398*	G/A	PCR1:	ACGTTGGATGAAGTCTTGTGATCCACCTGC
		PCR2:	ACGTTGGATGAAACACCATCAATCCCAGCC
		EXT:	CAAGCTCTCCTAGCTGAA
S243557_21170	C/T	PCR1:	ACGTTGGATGGCTTGCAACATTGCTATGGG
		PCR2:	ACGTTGGATGGAGAGGTCACATAACCTCTG
		EXT:	cccaCATCTATAGCTGCAGAAAAAA
S243769_14753	T/C	PCR1:	ACGTTGGATGCTGGTTTTAATGCCCTGTGG
		PCR2:	ACGTTGGATGAGTATAGTGGAAGACCCCAG
		EXT:	ccccTGTCCTGGCCCATCAT
S245852_10102	C/T	PCR1:	ACGTTGGATGAAAGGGATGTGACTGCGATG
		PCR2:	ACGTTGGATGCGGTGAAGTGTAATTTCTGC
		EXT:	ttcatAATTCTGCACTCTGGTA
S247283_18185	G/A	PCR1:	ACGTTGGATGTCTGAAGCTCTCCCACACTG
		PCR2:	ACGTTGGATGCTGTACATACTCCCAGCAGT
		EXT:	TCTGCAGTAAAGTGCG
S3038_262*	G/C	PCR1:	ACGTTGGATGTGCAGAGACACAGACAAACG
		PCR2:	ACGTTGGATGAGTCTTCTGGGAGGTTGTTC
		EXT:	AGAGGCACAGATTTAAACACA

* Excluded from SNP panels for genotyping

For those individuals collected from FBCC, additional microsatellite genotyping was conducted following the protocol described in Austin et al. (2012). Briefly, after DNA extraction and dilution (20 ng/µL), all samples were genotyped at 10 microsatellite loci published by Seyoum et al. (2013) (Table 2). Forward microsatellite primers were tagged with a 5’ fluorescent label [6-carboxyfluorescein (FAM), hexachlorofluorescein (HEX), or carboxytetramethylrhodamine (TAMRA)] to facilitate the use of multiplex PCR reactions. PCR products were electrophoresed using an ABI 3130xl (Applied Biosystems, Foster City, California) with a ROX 500 size standard (Applied Biosystems). Allele peaks were analyzed using GENEMARKER software (SoftGenetics, LLC, State College, PA) and all called alleles were manually confirmed.

Table 2 The 10 Microsatellite primers and fluorescent labels for the Florida bass. Each microsatellite primer was tagged with one of three fluorescent dyes [6-carboxyfluorescein (FAM), hexachlorofluorescein (HEX), carboxytetramethylrhodamine (TAMRA)]

Locus	Primer sequence	Repeat	Label	Genbank
Masf05	F: CGTCACCTCAGCCTCTGATT R: TCAGCAGCAACCAAAACAAC	(AC)^{14}	TAMRA	EU180167
Masf06	F: GACAGTGCACCAGGCCAAG R: ATCTGCAGGAGATTCTAGAGGATG	(AC)^{13}	FAM	EU180168
Masf10	F: ATCCCTCTCCCTCACTCTCTCTAT R: AAACTGTTTGAAATCTTTTGTTCCA	(CA)^{19}	HEX	EU180171
Masf17	F: AGGTTGCAGGAGCAGCAGCTAGAGCA R: ACGATGAGCCCTGTTGGGAGCTGT	(GT)^{18}	FAM	EU180175
Masf22	F: CCGAGCAGGGCAGCAGGAGAGGCAAG R: ACTTTATGTCTGAAGAGCAGTGACA	(CA)^{16}	FAM	EU180177
Masf24	F: CAGGCCCTTCCCCCATCCTTCCCCC R: TTGGCACGGGGAGGGAGACGAGTAT	(CT)^{20}	HEX	EU180163
Masf27	F: CTTCAGTTTAGCAGTTTACAGGGTTG R: ATGCAGCTCAAACTGATCCAC	(GT)^{41}	HEX	GU085830
Masf28	F: TCTTATGTTTCTGTTTTTAGGCATCA R: CTTTGGTCAGCTCTGTTCATACTCT	(CA)^{16}	FAM	GU085831
Masf29	F: CGTTCTCTGAAAATGTTTCACTTC R: ATACAATTTCTCACATTGTCTCTGTAG	(CA)^{23}	HEX	GU085832
Masf32	F: CCCCTTCATCAGATTTTATATGGTT R: AGGTCACATGCTGACTTTGTTACAC	(AC)^{13}	HEX	GU085834

2.4 Parentage Analysis

Two parentage analysis software packages, Cervus3 [16, 48] and SNPPIT 1.0 [39] were used for parentage assignment with and without sex information of parents. Parentage analyses in Cervus3 are mainly composed of three continuous modules: allele frequency analysis, simulation of parentage analysis and parentage analysis. First, the count and frequency of alleles at each genotyped locus were calculated. Second, Cervus3 determines the threshold log-likelihood (LOD) scores for the true parental pair by simulating the parents and offspring. I simulated 10,000 offspring for each tested population with default genotyping error rate 1% and default confidence levels (strict confidence 95% and relaxed confidence 80%). Simulations assumed 100% of parents sampled with candidate mothers and fathers set as 30 and 29, respectively. The proportion of loci genotyped was derived from allele frequency analysis. Lastly, the parentage analysis module assigns the most-likely candidate parent pair to each offspring tested with pre-determined population-wide assignment confidence. I only accepted parentage assignments with strict confidence (> 95%).

Another likelihood-based parentage assignment program, SNPPIT, was also used for assessing the assignment power of SNP panels developed from GBS data. Unlike Cervus3, SNPPIT was specifically designed for massive-scale parentage analysis of hatchery-released aquatic species using only biallelic SNP markers [39]. This program uses a likelihood-based categorical assignment method and Monte Carlo simulation to assess the confidence of the parentage assignment. Additionally, with the help of the pre-screening step based on Mendelian incompatibility, SNPPIT can dramatically reduce computation time. For assessing confidence in parentage assignments, SNPPIT uses false-discovery rates (FDR) instead of the population-wide assignment confidence used in Cervus3. To compare the results of parentage assignment between the two programs, I used a genotyping error rate of 1% or a per allele rate of 0.5% as a conservative parameter because it is larger than the actual per locus error rate [16]. Only parentage assignments with stringent FDR < 1% in SNPPIT were accepted.

While I initially used 71 SNPs for sample genotyping, 13 SNPs were subsequently dropped due to poor cluster distribution of genotypes and low genotype concordance (between GBS and MassARRAY), leaving 58 SNPs for further parentage analysis. The 250 GFEC samples comprised three known-pedigree families with two parent samples in each family. The FBCC population was composed of 53 adults and 211 juveniles collected from three ponds. For the FBCC population, as I only knew the potential adults that may have contributed to reproduction in each pond, two approaches were employed to assess the correctness of assignment. First, I compared the assignment to the spawning records and biological feasibility, that is, whether an offspring was assigned to the parent pair from wrong spawning ponds and whether the parent pair was of the opposite sex. Second, I employed a “gold standard” pedigree to assess assignment accuracy [21, 49] which required unambiguous assignment of parentage with known parent sex in Cervus3 based on pooled SNP (58) and microsatellite (10) genotypes at stringent confidence parameters (> 95%). To select an SNP panel with a sufficient number of loci to answer future management questions and to empirically test the assignment power of these selected loci, I assessed the assignment ability of different sized SNP sets. I first ranked all 58 SNPs based on their MAFs by pooling three Florida bass populations, computing MAFs for the pooled population, and then ranking the SNPs by their MAF. I then conducted parentage assignment using SNP sets with sequentially larger numbers of the top-ranked SNPs (i.e., SNPs with the highest MAF). To examine the resolving ability of SNPs, genotype data was pooled and analyzed. Considering that SNPPIT can only identify parent pairs with biallelic markers, assignment accuracy with SNPs was measured based on the correctness of parent pair assignments. To compare the assignment accuracy of the 58 SNP panel with 10 microsatellites, I conducted Cervus3 parent pair assignment with and without known parent sex for FBCC offspring. Assignment accuracy with two marker types was determined by comparing the results against the “gold standard”. To evaluate the assignment power in a more conservative scenario (i.e. a smaller proportion of parents sampled and high relatedness between candidate parents), I used varied proportions of the parents sampled (0.25, 0.50, 0.75, 1.00) and different relatedness levels among parents (0.25, 0.50, 0.75, 1.00) for simulation analysis in Cervus3 with 58 SNPs. For all assignments above, I additionally quantified the unassigned rate, the proportion of offspring that were not assigned to their parental pair despite the parental genotypes being present in the dataset, to evaluate the assignment ability of SNPs. The program ML-Relate [16] was used to calculate the likelihood of pedigree relationships (unrelated, half-sibs or full-sibs) for each pair of parental individuals and generate a matrix of relationships with the highest likelihood for that pair.

2.5. Statistical Analysis and Genotyping Error Estimate

Population summary statistics were calculated for all genotyped populations using SNPs and microsatellites. GenAlEx 6.5 [50, 51] was used to calculate the MAF, the observed (H_o) and expected (H_e) heterozygosity. Hardy-Weinberg equilibrium (HWE; exact test using a Markov Chain algorithm, with parameter settings of 1000 batches and 1000 iterations) was tested in each population using GENEPOP v. 4.5.1 [52]. Linkage disequilibrium (LD) between all pairs of loci was analyzed for adult fish in each population using a simulated exact test in GENEPOP. The false discovery rate (FDR) correction was used for both HWE and LD tests to reduce the type I errors. For microsatellites alone, Micro-Checker software version 2.2.3 [53] was used to test for the presence of scoring artifacts in the brood fish of FBCC.

All types of molecular markers are prone to genotyping errors and a high error rate can lead to false parentage assignment and consequently bias biological conclusions [54]. Therefore, I used 15 individuals from the STJR population to test the concordance of SNP genotypes (58 SNPs) generated with MassARRAY and GBS. Genotyping error rate of 58 SNPs for each method was then estimated by comparing the genotyping of individuals that were taken through all steps at least twice (DNA extraction of tissues through genotyping protocols), which were performed using 19 individuals from FBCC for MassARRAY System (two replicates for each individual) and six individuals from GFEC for the GBS (four replicates). Discordant genotypes due to missing data were excluded in this analysis.

3. Results and Discussion

3.1 SNP Marker Discovery and Genotyping

The dramatic decline in next-generation sequencing (NGS) costs has increased the accessibility of SNP markers for population genetics and genomics in nonmodel organisms [55, 56]. GBS represents a cost-effective tool that allows high-throughput and simultaneous SNP discovery and genotyping for target organisms [23]. SNPs, with the advantages of high transferability, high-throughput, and low genotyping error, are expected to replace microsatellites for parentage determination in aquatic species [19, 22]. Herein, I utilized 265 hatchery-reared Florida bass for initial GBS analysis (GFEC and STJR) and generated a total of 634,226,037 high-quality reads from three Illumina HiSeq lanes with an average of 2,233,190 reads for each sequenced sample. The TASSEL-GBS pipeline clustered reads into 2,400,888 locus-specific tags. During the BWA alignment to the largemouth bass draft genome, 1,561,215 tags were aligned to unique positions (65.03%), 616,038 tags were aligned to multiple positions (25.66%), and 223,635 tags could not be aligned (9.31%). I carried forward the unique mapped tags for SNP discovery. A total of 58,450 pre-filtered SNPs were generated from TASSEL package. During this step, to call a SNP within an individual, a minimum MAF of 0.05, a minimum LCov of 0.1, and a minimum MAC of 10 criteria had to be met. Further stringent filtering processes (criteria 1–6; Methods) conducted using VCFtools resulted in the identification of 250 biallelic polymorphic loci with MAF > 0.4 for subsequent multiplex parentage panel design (Fig. 1).

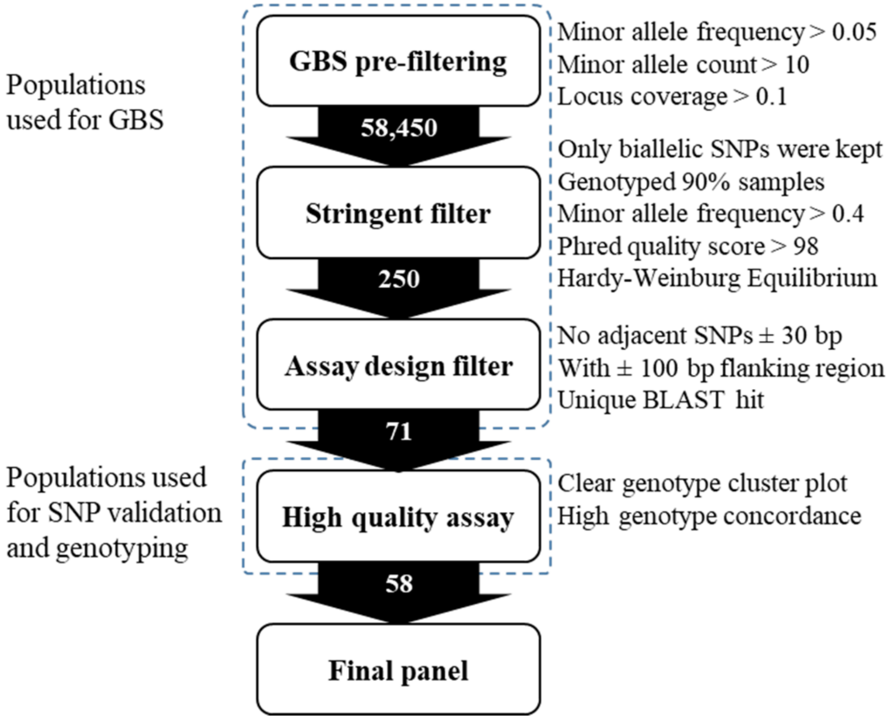


Fig.1 The workflow outlining the steps used in marker identification and selection of SNP panels for parentage assignment.

While GBS provides a cost-effective and rapid means for initial SNP discovery and genotyping, the practical application of SNP markers in parentage assignment of Florida bass requires the use of smaller, flexible multiplex panels for genotyping [41]. Given previous parentage modeling and experiments in fish [8, 21, 57], I initially chose 100 SNP loci from the above set of 250 for Agena MassARRAY assay design. Of these, 71 SNPs were successfully multiplexed and used for MassARRAY genotyping of 661 *Micropterus* individuals collected from six populations, both to confirm the GBS genotypes for SNPs selected based on stringent filtering parameters as well as to assess marker polymorphism and MAF across diverse populations. The six populations are listed in Table 3, and included both GBS samples (GFEC and STJR) and additional geographically distributed samples with varying degrees of hybridization/introgression with *M. salmoides*. All SNPs in two multiplexes were successfully amplified in both GBS populations and additional bass populations of varying degrees of introgression with *M. salmoides*. Of the 71 SNPs, 58 (81.69%) performed well based on genotype cluster distribution and concordance. The poor quality genotyping plots and low concordance of the remaining SNPs were potentially due to the paralogous loci unidentified based on BLAST alignment against the draft genome [41]. Comparison of the SNP genotypes derived from GBS and MassARRAY data using 15 STJR individuals revealed a high concordance (99.7%) with only three genotype discrepancies. Comparison of SNP genotypes from replicated samples yielded a mean error rate of 0.6% (8 discrepancies out of 1392 data comparisons) and 0.0% (2204 data comparisons) for the GBS and MassARRAY chemistries, respectively. The results confirmed that the MassARRAY system is a highly sensitive and accurate method for SNP detection and validation as previously reported by Oeth et al. (2005) [58].

Table 3 Summary statistics of 58 SNP loci across six populations. The introgression level in each population was assessed using diagnostic SNP panels previously developed in our lab. Summary statistics of 10 microsatellites (Micro) are listed in the second row. Population abbreviations are: FBCC is Florida Bass Conservation Center, FL; STJR is St. Johns River, FL; GFEC is GO Fish Education Center, GA; ALLA is Lake Allatoona, GA; DKLA is Dekalb County Lake, AL; OLHR is Old Hickory Reservoir, TN.

Population	*N*	*P*	MAF	H_o	H_{e}	*M. salmoides*	*M. floridanus*
FBCC (SNPs)	264	93.1	0.26	0.37	0.3	0.01	0.99
FBCC (Micro)	264	100.0	NA	0.73	0.7		
STJR	15	93.1	0.26	0.38	0.3	0.01	0.99
GFEC	250	100.0	0.45	0.51	0.5	0.04	0.96
ALLA	100	100.0	0.27	0.34	0.4	0.50	0.50
DKLA	16	65.5	0.08	0.13	0.1	0.84	0.16
OLHR	16	43.1	0.05	0.08	0.1	0.94	0.06

P = percentage of polymorphic loci, H_o = average observed heterozygosity, H_{e} = average expected heterozygosity, MAF = average minor allele frequency, NA = not applicable.

3.2 Marker Characteristics and MAF Distributions

High minor allele frequency (MAF) in populations has been shown to be one of the most important criteria for guiding selection of informative SNPs for parentage assignment [8, 57]. I measured the MAF distribution in each of six populations (Fig. 2; Table 3). Each of the 58 SNPs was polymorphic in multiple populations. The MAF distribution was skewed toward higher MAF in three Florida bass-derived populations (FBCC, GFEC and STJR, Table 3), with mean MAF ranging from 0.26 (STJR and FBCC) to 0.45 (GFEC). The mean MAF across these three populations was 0.32. A similar MAF distribution (mean MAF 0.34) was recently reported in parentage analysis for rainbow trout (*Oncorhynchus mykiss*) using a panel of 95 SNP assays [8]. Perfect parentage assignments obtained with 95 SNPs in that study indicated that SNPs with high MAF were a promising predictor for parentage success [8].

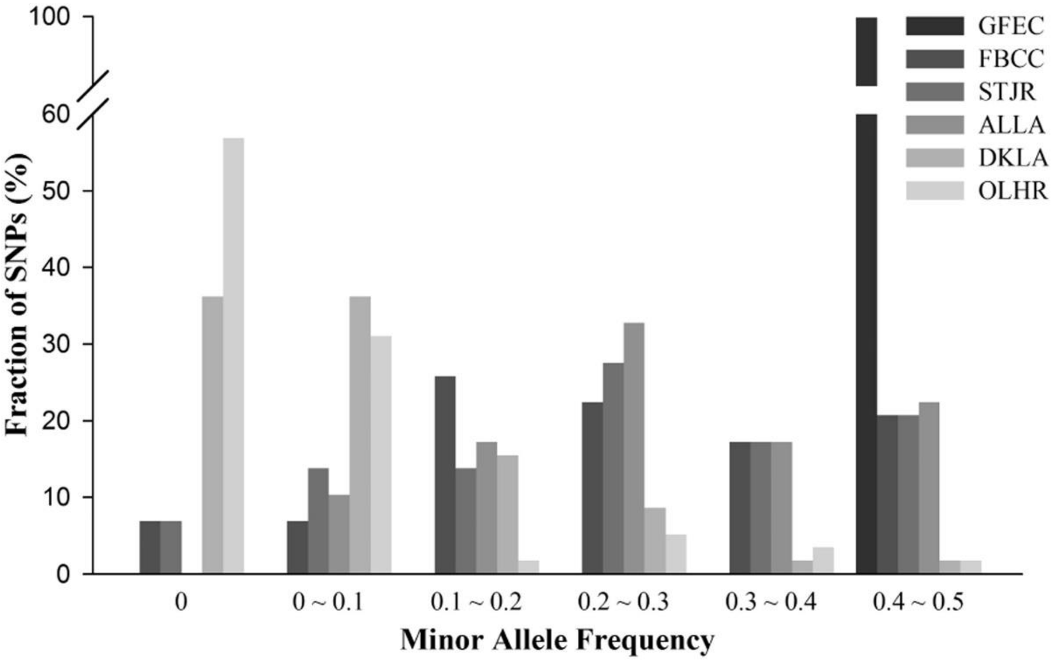


Fig. 2 Minor allele frequency distributions for 58 SNPs across six populations. Population abbreviations are: FBCC is Florida Bass Conservation Center, FL; STJR is St. Johns River, FL; GFEC is GO Fish Education Center, GA; ALLA is Lake Allatoona, GA; DKLA is Dekalb County Lake, AL; OLHR is Old Hickory Reservoir, TN.

Florida bass readily hybridize with their sister taxa, *M. salmoides,* when stocked together into the same waters [34]. Additionally, a natural intergrade zone existed prior to the onset of widespread stocking [59, 60]. Outside of peninsular Florida, therefore, genetic analyses necessarily involve both species and their hybrid. I was therefore interested in the MAF distribution of these markers in populations with increasing *M. salmoides* allele frequencies. MAF remained high in samples from Lake Allatoona, GA, which had equal allelic contributions from *M. floridanus* and *M. salmoides* (mean MAF of 0.27; Table 3). In contrast, samples from Dekalb County Lake, AL which were 84% *M. salmoides,* showed a mean MAF of 0.08 (range from 0.00 to 0.44), with 20 SNPs (34%) showing monomorphism. Similarly, SNPs screened on individual fish collected from Old Hickory Reservoir, TN (94% *M. salmoides*) also exhibited a relatively low mean MAF of 0.05 (ranging from 0.00 to 0.41), with 33 SNPs (56.90%) being monomorphic. A recent study using 60 SNP makers also showed marked MAF distinctions among pure *Mytilus trossulus*, pure *Mytilus* edulis and an introgressed population composed of *M. edulis* − *M. trossulus* hybrids [61]. Because the multiplex assays were developed from "pure" *M. floridanus* samples, it was not surprising that marker polymorphism within Florida bass populations was higher than within *M. salmoides*-dominated populations. The higher average MAF in a hybrid population (ALLA), however, demonstrated the likelihood that parentage analysis could be conducted in introgressed populations (with ~ 50% *M. floridanus* alleles) using the 58-SNP panel.

Further population statistics are summarized in Table 3. For the 58 SNPs, the average observed heterozygosity (H_o) over all loci varied from 0.08 (OLHR) to 0.51 (GFEC), while the average expected heterozygosity ranged from 0.08 in OLHR to 0.49 in GFEC. Examination of HWE deviations in the six populations involved 348 tests for heterozygote deficiency with only one locus showing significant deviation after FDR correction. Of the 9918 linkage disequilibrium

(LD) tests in 58 SNPs, there was one significant locus pair following FDR correction. I also calculated population-level metrics for 10 microsatellites (Table S4) used in subsequent parentage analysis comparisons. For the 10 microsatellites, the mean number of alleles per locus was 10.4 in FBCC parental samples. The mean observed and expected heterozygosities were 0.73 and 0.72, respectively (Table 3). Of the 10 microsatellites, one locus deviated significantly from HWE, while no loci showed evidence for LD or null alleles (data not shown). Based on these results, all 58 SNPs and 10 microsatellites were retained for downstream parentage analysis.

3.3 Accuracy of Parentage Assignment Using SNPs and Microsatellites

Before parentage assignment evaluation, I examined pedigree information in the GFEC population used for training (3 full-sib families with 244 offspring with known parentage). Genotypes from the 58 SNPs analyzed through either Cervus3 or SNPPIT resulted in parentage assignment with 100% accuracy. For FBCC test individuals (211 offspring, 53 parents), I relied on a "gold standard" [21] for assessing assignment accuracy. This "gold standard" (against which SNP alone and microsatellite alone assignments would be compared) required unambiguous assignment of parentage in Cervus3 based on pooled SNP (58) and microsatellite (10) genotypes at stringent confidence parameters (> 95%). I also checked all assignments for plausibility based on spawning records and biological feasibility. The "gold standard" from Cervus3 (58 SNPs plus 10 microsatellites) resulted in matching assignments of 211 offspring to 13 broodfish. I detected evidence of monogamy, consistent with the mating patterns recently reported in Florida bass [37, 38]. Based on assignment results, two males (15.4% of identified parents) and one female (7.7%) spawned more than once, which was consistent with Isaac et al. who noted multiple spawning times of Florida bass [62].

I then compared the “gold standard” assignments derived from both marker types with those generated using SNP or microsatellite data alone. The 58 SNP panels alone resulted in assignments matching the “gold standard” using Cervus3 (> 95% confidence) in all cases (211 offspring with 422 parent assignments), irrespective of whether the sex of the parents was used in the analyses (Fig. 3). On the other hand, the 10 microsatellites, when used alone, resulted in a small number (13) of misassigned parents when sex was known, and a larger number of offspring (49) with unassigned parents when sex was unknown (not used in analyses; Fig. 3). Microsatellites have been widely used for parentage studies in aquatic animals because of their high variability and wide availability [21]. It has been estimated that approximately six SNPs have equal assignment power to one microsatellite [63]. The comparison here indicates that the selected SNPs may be more informative than would be predicted by this generalized ratio. Combined with comparative advantages in multiplexing, ease of scoring, reproducibility, and lower associated labor costs, there appears to be a solid rationale for transitioning bass parentage approaches toward SNP platforms.

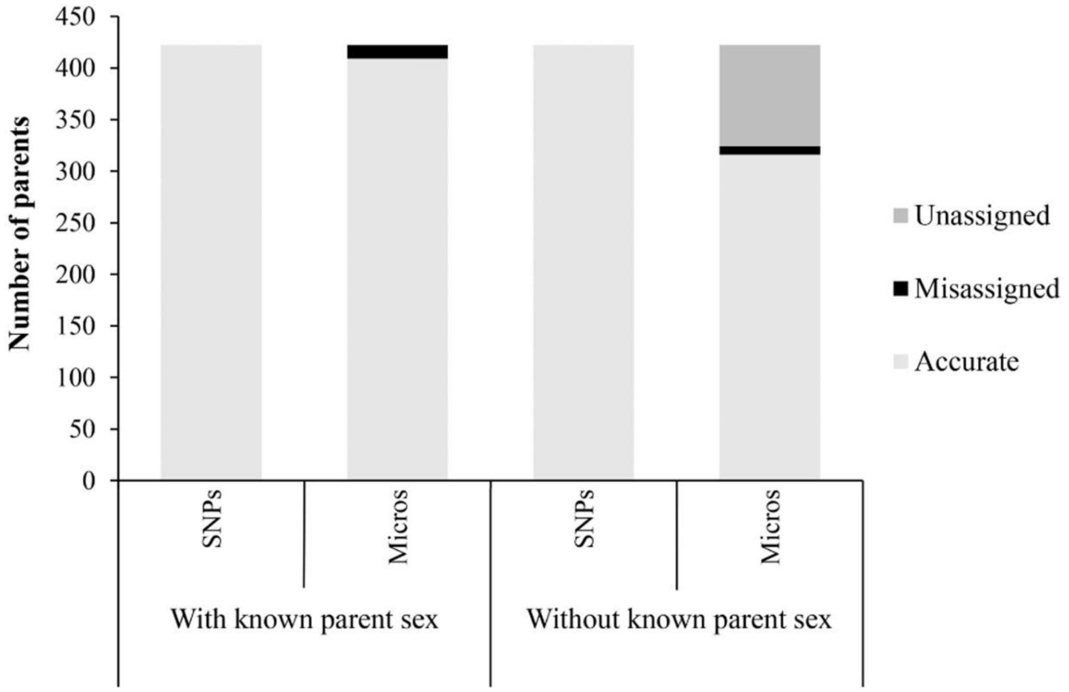


Fig. 3 Parent pair assignments compared to the "gold standard" for 211 FBCC offspring (i.e. 422 parents) using SNP and microsatellite markers. Bar sections from bottom to top: identical assignment between "gold standard" and respective method (Accurate, assignment confidence > 95%), different parents assigned by the gold standard and respective method (Mis-assigned, assignment confidence > 95%), assigned by gold standard but not respective method (Unassigned, assignment confidence < 95%). Gold standard unambiguous assignment of parentage with known parent sex in Cervus3 based on pooled SNP (58) and microsatellite (10) genotypes at stringent confidence parameters (> 95%).

Following the direct comparison of the two marker types, I utilized the 58 SNP genotypes from both the GFEC and FBCC datasets to compare two commonly used parentage assignment packages, Cervus3, and SNPPIT. Based on assignment accuracy and the number of high confidence assignments, SNPPIT outperformed Cervus3 (Fig. 4 and Table 4). Parentage assignment using the full 58 SNPs performed equally well in both SNPPIT and Cervus3, with a 100% parentage accuracy and 0 unassigned individual fish. I then ranked the 58 SNPs based on MAF and tested the most informative subset of 28, 38, and 48 SNPs to assess the power of smaller groups of SNPs for parentage analysis using the two programs. As seen in Table 4 and Fig. 4, using 28 SNPs and SNPPIT analyses resulted in > 90% accurate assignments when sex was known and > 80% when sex was not known. These metrics differed substantially from Cervus3 which only assigned ~ 60% of offspring accurately with sex and ~ 4% without, at the level of 28 SNPs. The low rate of accurate assignment using 28 SNPs in Cervus3 was likely due to differences in how confidence is assessed. In Cervus3, the stated confidence level is the average confidence in all of the accepted assignments, and Cervus3 will therefore not assign parents to some individuals where the confidence in that particular assignment is lower than the stated population-level confidence [49]. When 28 SNPs were used for parentage assignment with known sex, the stated population-level confidence was higher than that in 27.87% of the GFEC assignments, resulting in a relatively high unassigned rate. However, with at least 38 SNPs, differences in assignment accuracy between the two programs, irrespective of parent sex data, became minimal.

Table 4 Percentage of samples that were accurately assigned, misassigned, or were unassigned using different numbers of SNPs for parentage analysis with and without parental sex information. SNPs for the differently sized sets were ranked and selected based on the MAFs within three Florida bass populations. "Unassigned" represents the portion of offspring known to have parents represented in the dataset but were not assigned regardless. GFEC, GO Fish Education Center, GA; FBCC, Florida Bass Conservation Center, FL.

No. of loci		% Accurate	% Misassignment	% Unassigned
		With known sex		
GFEC	28	90.57/58.61	7.38/13.52	2.05/27.87
	38	96.72/95.49	3.28/4.51	0.00/0.00
	48	96.72/97.13	3.28/2.87	0.00/0.00
	58	100.00/100.00	0.00/0.00	0.00/0.00
FBCC	28	97.63/94.31	1.90/2.84	0.47/2.84
	38	100.0/100.0	0.00/0.00	0.00/0.00
	48	100.0/100.0	0.00/0.00	0.00/0.00
	58	100.0/100.0	0.00/0.00	0.00/0.00
		Without known sex		
GFEC	28	81.15/3.69	10.24/2.05	8.61/94.26
	38	96.72/93.03	3.28/6.97	0.00/0.00
	48	96.72/97.13	3.28/2.87	0.00/0.00
	58	100.00/100.00	0.00/0.00	0.00/0.00

FBCC	28	96.21/50.71	3.32/1.42	0.47/47.87
	38	100.0/100.0	0.00/0.00	0.00/0.00
	48	100.0/100.0	0.00/0.00	0.00/0.00
	58	100.0/100.0	0.00/0.00	0.00/0.00

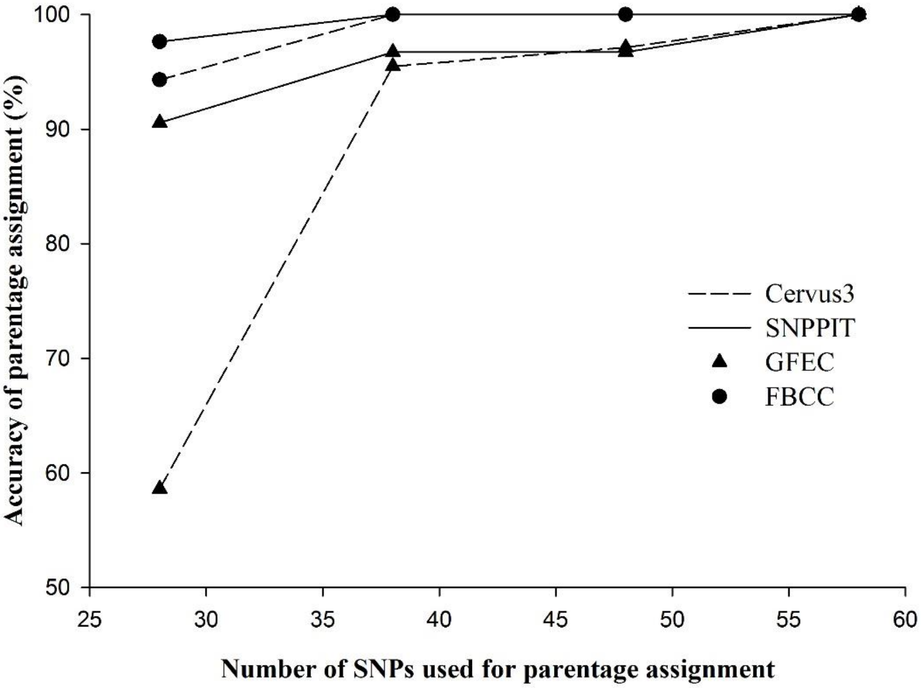


Fig. 4 The accuracy of parentage assignment with known parent sex in FBCC and GFEC populations using differently sized SNP sets. SNPs for the differently sized sets were ranked and selected based on the MAFs within three Florida bass populations. FBCC, Florida Bass Conservation Center, FL; GFEC, GO Fish Education Center, GA.

Several factors may affect the accuracy of parentage assignment. One of the main factors is the presence of kinship in the samples [64]. Generally, a nonparent relative of either the offspring or a true parent is likely to be misclassified as a parent [64]. A previous study using microsatellites for chinook salmon pedigree reconstruction concluded that the presence of full-sib candidate parents could substantially decrease the resolving power of genetic markers for parentage analysis [65]. In this study, a pair of full-sib adults was identified in the GFEC population, based on ML-Relate kinship analysis, while adult pairs in FBCC showed no evidence for full-sib relationships

(Fig. 5), likely explaining, at least in part, differences in assignment performance between GFEC and FBCC populations. Nonetheless, in both populations, 100% assignment accuracy was obtained when the 58-SNP panel was used, indicating the sufficient power of these SNPs to resolve pedigrees with full-sib candidate parents. Additionally, simulation analyses in Cervus3 using varied proportions of the parents sampled (0.25, 0.50, 0.75, 1.00) and different relatedness levels among parents (0.25, 0.50, 0.75, 1.00) both achieved 100% accuracy (Table 5), demonstrating the sufficient power of 58 SNPs in parentage assignment when a small proportion of parents were sampled and/or the presence of high relatedness among candidate parents.

Table 5 The accuracy of parentage assignment in simulation studies using varied proportions of the parents sampled and different relatedness levels among true parents. Assignments were performed using FBCC and GFEC populations with 58 SNPs. The relatedness of true parents is the probability that an allele in one parent is identical by descent to the corresponding allele in the other parent.

	% Accurate	% Misassignment	%
Proportion of parents sampled		With known sex	
1.00	100.00	0.00	0.00
0.75	100.00	0.00	0.00
0.50	100.00	0.00	0.00
0.25	100.00	0.00	0.00
		Without known sex	
1.00	100.00	0.00	0.00
0.75	100.00	0.00	0.00
0.50	100.00	0.00	0.00
0.25	100.00	0.00	0.00
Relatedness of true parents		With known sex	
1.00	100.00	0.00	0.00
0.75	100.00	0.00	0.00
0.50	100.00	0.00	0.00
0.25	100.00	0.00	0.00
		Without known sex	
1.00	100.00	0.00	0.00
0.75	100.00	0.00	0.00
0.50	100.00	0.00	0.00
0.25	100.00	0.00	0.00

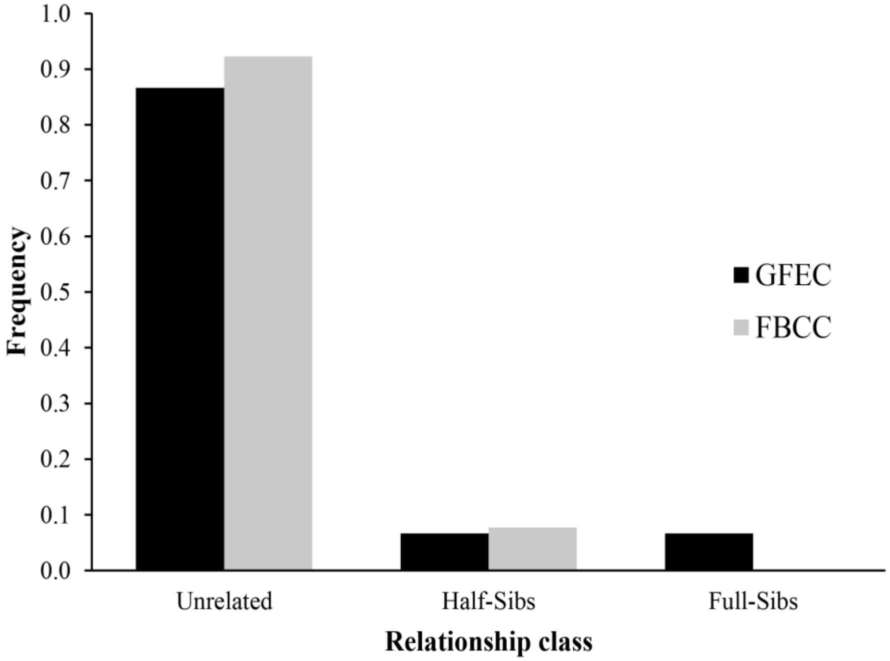


Fig. 5 Frequency distribution of maximum likelihood relatedness estimates for pairs of parental individuals (in GFEC and FBCC populations) classified as unrelated, half-sibs or full-sibs using the program ML-Relate. GFEC, GO Fish Education Center, GA; FBCC, Florida Bass Conservation Center, FL.

Although SNPs offer several advantages over microsatellites, the historically high cost of multiplex SNP genotyping and the necessity to test a higher number of SNPs to select optimal ones have limited their wide application in parentage analysis [14]. A small proportion of filtered SNPs had to be excluded from further parentage analysis due to their poor performance in genotyping clustering and/or a lack of concordance between platforms. Initial mining of polymorphic SNPs (Fig. 1) from GBS sequencing of the three GFEC families (part of an additional, ongoing linkage mapping project) indisputably biased selected markers toward informativeness in this population. However, I demonstrate here that these markers remain polymorphic in a number of pure and intergrade populations and are of high utility for parentage analysis in unrelated *M. floridanus* populations (e.g., FBCC with MAF of 0.26). Future testing and validation in additional wild and hatchery samples with known parentage relationships should allow further refinement of the current marker panels. Additionally, although SNPs used in this study showed little evidence of linkage disequilibrium (LD), future marker mapping with the benefit of linkage groups should ensure even distribution of these and future markers across the bass genome.

4. Conclusions

The development of SNP markers using next-generation sequencing platforms such as GBS represents a cost-effective and efficient means to identify parent-offspring pairs and assess population genetic patterns. Given the global distribution of black bass (*Micropterus* spp.; Hargrove et al., 2015) and high frequency with which bass are stocked (Heidinger, 1976) developing quantitative tools to assess ancestry (Li et al., 2015) and relatedness are of utmost significance to propagation, management, and conservation needs. In this study, I described a SNP set for Florida bass that is capable of identifying parent-offspring pairs with very high accuracy. Furthermore, I validated these SNPs relative to microsatellite markers using a "gold standard" pedigree and identify reduced SNP sets that can be used to perform parentage analysis in introgressed populations. Taken together, these markers represent a valuable tool for a myriad of downstream applications in genetic management, breeding, and stock enhancement in Florida bass.

References

1. Allendorf, F., Genetic management of hatchery stocks. Population Genetics and Fishery Management, 1987: p. 141-159.
2. Bert, T.M., et al., Genetic management of hatchery-based stock enhancement, in Ecological and genetic implications of aquaculture activities. 2007, Springer. p. 123-174.
3. MacEina, M.J. and B.R. Murphy, Stocking Florida largemouth bass outside its native range. 1992.
4. Sekino, M., et al., Genetic tagging of released Japanese flounder (*Paralichthys olivaceus*) based on polymorphic DNA markers. Aquaculture, 2005. **244**(1-4): p. 49-61.
5. Waples, R.S. and J. Drake, Risk/benefit considerations for marine stock enhancement: a Pacific salmon perspective. Stock Enhancement and Sea Ranching. Developments Pitfalls and Opportunities, 2004: p. 260-306.
6. Buynak, G.L. and B. Mitchell, Contribution of stocked advanced-fingerling largemouth bass to the population and fishery at Taylorsville Lake, Kentucky. North American Journal of Fisheries Management, 1999. **19**(2): p. 494-503.
7. Buynak, G.L., et al., Stocking subadult largemouth bass to meet angler expectations at Carr Creek Lake, Kentucky. North American Journal of Fisheries Management, 1999. **19**(4): p. 1017-1027.
8. Liu, S., et al., Development and validation of a SNP panel for parentage assignment in rainbow trout. Aquaculture, 2016. **452**: p. 178-182.
9. Bergman, P.K., et al., Perspectives on design, use, and misuse of fish tags. Fisheries, 1992. **17**(4): p. 20-25.

10. Steele, C.A., et al., A validation of parentage-based tagging using hatchery steelhead in the Snake River basin. Canadian Journal of Fisheries and Aquatic Sciences, 2013. **70**(7): p. 1046-1054.

11. Jepsen, N., et al., The use of external electronic tags on fish: an evaluation of tag retention and tagging effects. Animal Biotelemetry, 2015. **3**(1): p. 49.

12. Estoup, A., et al., Parentage assignment using microsatellites in turbot (Scophthalmus maximus) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. Canadian Journal of Fisheries and Aquatic Sciences, 1998. **55**(3): p. 715-723.

13. Herbinger, C.M., et al., DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture, 1995. **137**(1-4): p. 245-256.

14. Vandeputte, M. and P. Haffray, Parentage assignment with genomic markers: a major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. Frontiers in Genetics, 2014. **5**: p. 432.

15. Ball, A.D., et al., A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). BMC Genomics, 2010. **11**(1): p. 218.

16. Kalinowski, S.T., M.L. Taper, and T.C. Marshall, Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology, 2007. **16**(5): p. 1099-1106.

17. Pritchard, V., A. Abadía‐Cardoso, and J. Garza, Discovery and characterization of a large number of diagnostic markers to discriminate Oncorhynchus mykiss and O. clarkii. Molecular Ecology Resources, 2012. **12**(5): p. 918-931.

18. Slate, J., et al., Gene mapping in the wild with SNPs: guidelines and future directions. Genetica, 2009. **136**(1): p. 97-107.

19. Jin, Y.L., et al., Development, inheritance and evaluation of 55 novel single nucleotide polymorphism markers for parentage assignment in the Pacific oyster (*Crassostrea gigas*). Genes & Genomics, 2014. **36**(2): p. 129-141.

20. Hess, J.E., et al., Use of genotyping by sequencing data to develop a high-throughput and multifunctional SNP panel for conservation applications in Pacific lamprey. Molecular Ecology Resources, 2015. **15**(1): p. 187-202.

21. Hauser, L., et al., An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. Molecular Ecology Resources, 2011. **11**: p. 150-161.

22. Yue, G.H. and J.H. Xia, Practical considerations of molecular parentage analysis in fish. Journal of the World Aquaculture Society, 2014. **45**(2): p. 89-103.

23. Kaiser, S.A., et al., A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. Molecular Ecology Resources, 2017. **17**(2): p. 183-193.

24. Elshire, R.J., et al., A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One, 2011. **6**(5): p. e19379.

25. Uncu, A.O., et al., High-throughput single nucleotide polymorphism (SNP) identification and mapping in the sesame (*Sesamum indicum L.*) genome with genotyping by sequencing (GBS) analysis. Molecular Breeding, 2016. **36**(12): p. 173.

26. Bielenberg, D.G., et al., Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) Batsch]. PLoS One, 2015. **10**(10): p. e0139406.

27. Kim, C., et al., Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Science, 2016. **242**: p. 14-22.

28. Liu, H., et al., An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. BMC Genomics, 2014. **15**(1): p. 104.

29. Peterson, G., et al., Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. Diversity, 2014. **6**(4): p. 665-680.

30. Barthel, B.L., et al., Genetic relationships among populations of Florida Bass. Transactions of the American Fisheries Society, 2010. **139**(6): p. 1615-1641.

31. Johnson, R. and T. Fulton, Persistence of Florida largemouth bass alleles in a northern Arkansas population of largemouth bass, *Micropterus salmoides Lacepede*. Ecology of Freshwater Fish, 1999. **8**(1): p. 35-42.

32. Li, C., et al., Discovery and validation of gene-linked diagnostic SNP markers for assessing hybridization between Largemouth bass (*Micropterus salmoides*) and Florida bass (*M. floridanus*). Molecular Ecology Resources, 2015. **15**(2): p. 395-404.

33. Philipp, D.P., W.F. Childers, and G.S. Whitt, A biochemical genetic evaluation of the northern and Florida subspecies of largemouth bass. Transactions of the American Fisheries Society, 1983. **112**(1): p. 1-20.

34. Maceina, M.J., B.R. Murphy, and J.J. Isely, Factors regulating Florida largemouth bass stocking success and hybridization with northern largemouth bass in Aquilla Lake, Texas. Transactions of the American Fisheries Society, 1988. **117**(3): p. 221-231.

35. Seyoum, S., et al., Isolation and characterization of eighteen microsatellite loci for the largemouth bass, *Micropterus salmoides*, and cross amplification in congeneric species. Conservation Genetics Resources, 2013. **5**(3): p. 697-701.

36. Hargrove, J.S., O.L. Weyl, and J.D. Austin, Reconstructing the introduction history of an invasive fish predator in South Africa. Biological Invasions, 2017. **19**(8): p. 2261-2276.

37. Austin, J.D., et al., An assessment of hatchery effects on Florida bass (*Micropterus salmoides floridanus*) microsatellite genetic diversity and sib‐ship reconstruction. Aquaculture Research, 2012. **43**(4): p. 628-638.

38. Hargrove, J.S. and J.D. Austin, Parentage and mating patterns in a Florida Largemouth Bass (*Micropterus salmoides floridanus*) hatchery. Aquaculture Research, 2017. **48**(6): p. 3272-3277.

39. Anderson, E., Computational algorithms and user-friendly software for parentage-based tagging of Pacific salmonids. Final report submitted to the Pacific Salmon Commission's Chinook Technical Committee (US Section), 2010.

40. De Donato, M., et al., Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One, 2013. **8**(5): p. e62137.

41. Li, C., et al., SNP discovery in wild and domesticated populations of blue catfish, *Ictalurus furcatus*, using genotyping‐by‐sequencing and subsequent SNP validation. Molecular Ecology Resources, 2014. **14**(6): p. 1261-1270.

42. Glaubitz, J.C., et al., TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One, 2014. **9**(2): p. e90346.

43. Zimin, A.V., et al., The MaSuRCA genome assembler. Bioinformatics, 2013. **29**(21): p. 2669-2677.

44. Ryan, J., estimate_genome_size. pl (Version 0.03)[Computer Software]. Sars International Centre for Marine Molecular Biology, Bergen, Norway (Retrieved from http://josephryan. github. com/estimate_genome_size. pl/), 2013.

45. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009. **25**(14): p. 1754-1760.

46. Danecek, P., et al., The variant call format and VCFtools. Bioinformatics, 2011. **27**(15): p. 2156-2158.

47. Johnson, J.L., et al., Genotyping-by-sequencing (GBS) detects genetic structure and confirms behavioral QTL in tame and aggressive foxes (*Vulpes vulpes*). PLoS One, 2015. **10**(6): p. e0127013.

48. Marshall, T., et al., Statistical confidence for likelihood‐based paternity inference in natural populations. Molecular Ecology, 1998. **7**(5): p. 639-655.

49. Walling, C.A., et al., Comparing parentage inference software: reanalysis of a red deer pedigree. Molecular Ecology, 2010. **19**(9): p. 1914-1928.

50. Smouse, R.P.P. and R. Peakall, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics, 2012. **28**(19): p. 2537-2539.

51. Peakall, R. and P.E. Smouse, GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes, 2006. **6**(1): p. 288-295.

52. Raymond, M. and F. Rousset, GENEPOP a population genetics software for exact tests and ecumenicism, version 2. Journal of Heredity, 1995. **86**: p. 248-269.

53. Van Oosterhout, C., et al., MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes, 2004. **4**(3): p. 535-538.

54. Pompanon, F., et al., Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics, 2005. **6**(11): p. 847.

55. Rice, A.M., et al., A guide to the genomics of ecological speciation in natural animal populations. Ecology Letters, 2011. **14**(1): p. 9-18.

56. Hohenlohe, P.A., et al., Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology Resources, 2011. **11**: p. 117-122.

57. Anderson, E.C. and J.C. Garza, The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics, 2006. **172**(4): p. 2567-2582.

58. Oeth, P., et al., iPLEX assay: Increased plexing efficiency and flexibility for MassArray system through single base primer extension with mass-modified terminators. Sequenom Application Note, 2005. **27**.

59. Near, T.J., et al., Speciation in North American black basses, *Micropterus* (*Actinopterygii: Centrarchidae*). Evolution, 2003. **57**(7): p. 1610-1621.

60. Nedbal, M.A. and D.P. Philipp, Differentiation of mitochondrial DNA in largemouth bass. Transactions of the American Fisheries Society, 1994. **123**(4): p. 460-468.

61. Zbawicka, M., et al., New SNP markers reveal largely concordant clinal variation across the hybrid zone between Mytilus spp. in the Baltic Sea. Aquatic Biology, 2014. **21**(1): p. 25-36.

62. Isaac Jr, J., et al., Spawning behavior of Florida largemouth bass in an indoor raceway. The Progressive Fish-Culturist, 1998. **60**(1): p. 59-62.

63. Glaubitz, J.C., O.E. Rhodes Jr, and J.A. DeWoody, Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Mol Ecol, 2003. **12**(4): p. 1039-1047.

64. Städele, V. and L. Vigilant, Strategies for determining kinship in wild populations using genetic data. Ecology and Evolution, 2016. **6**(17): p. 6107-6120.

65. Olsen, J., et al., The aunt and uncle effect: an empirical evaluation of the confounding influence of full sibs of parents on pedigree reconstruction. Journal of Heredity, 2001. **92**(3): p. 243-247.

Chapter III SNP panel development for genetic management of wild and domesticated white bass (*Morone chrysops*)

Abstract

White bass (*Morone chrysops*), striped bass and their interspecific hybrid are important game fishes, whereas the hybrid striped bass is an important aquaculture species in the US. Numerous states, federal and private hatcheries, therefore, rear these species for stocking purposes as well as for food fish. Although striped bass populations (both wild and domesticated) have been extensively evaluated, relatively little effort has been directed toward the study and improvement of white bass. In this study, I developed SNP resources to examine the genetic relationships among a long-term domesticated white bass line and five potential founder stocks for selective breeding collected from drainages in Arkansas, Texas, and Alabama. Using genotyping-by-sequencing, I generated 13,872 genome-wide SNP loci across the six populations. Stringent filtering of SNP-calling parameters identified 426 informative SNP loci. Population genetics and structure analyses using these loci revealed only moderate genetic differentiation between populations (global F_{st} = 0.083) and indicated two major genetic clusters. A final 57-SNP assay was successfully designed and validated using the MassARRAY system. The developed SNP panel assigned 96 additional genotyped individuals to their population of origin with 100% accuracy. The SNP resources developed in this study should facilitate ongoing efforts in selective breeding and conservation of white bass.

1. Introduction

White bass (*Morone chrysops*) is an economically valuable and ecologically important fish species in the US [1, 2]. Together with its sister species, striped bass (*Morone saxatilis*), white bass are prized for recreational angling, whereas the hybrid striped bass (*M. chrysops* × *M. saxatilis*) is a major commodity in US aquaculture production [3-5]. These species are reared by numerous states, federal and private hatcheries for stocking as well as food fish production. Although the genetic relationships and performance of striped bass wild populations and domestic strains have been fairly extensively characterized [6, 7], knowledge of genetic structure in white bass is minimal. This is due, in large part, to a paucity of modern, verified genetic markers for the species.

White bass are native to the Mississippi River drainage and lower Great Lakes but have been introduced widely outside this range, potentially clouding historical patterns of population structure. White (2000) [2] used protein electrophoresis at seven loci to examine genetic variation in Ohio River drainage white bass populations, noting exceptionally low heterozygosity and concluding that he could not detect ‘genetically meaningful stock structure outside of Lake Erie.’ More recently, Couch et al. (2006) [8] identified microsatellites from striped and white bass. However, only six white bass individuals were tested using these markers and the authors noted an average of only 2.2 alleles per locus.

The US hybrid striped bass industry currently sources wild white bass broodstock from river systems across the eastern US. Genetic and genomic tools are needed to support the ongoing domestication and selective breeding efforts of the USDA-ARS Stuttgart National Aquaculture Research Center (SNARC; Stuttgart, AR), aimed at providing the industry with a superior and sustainable, cultured broodstock source [3, 9-11]. Given the complex stocking history of white bass and suspected low genetic structuring [2, 4], sourcing wild founder stocks from widespread geographical localities does not necessarily ensure genetic diversity. The study goal here, therefore, was to develop and validate single nucleotide polymorphism (SNP) markers with high utility for: (i) resolving genetic variation among potential founder stocks and (ii) accurately assigning individuals to source populations to aid in long-term broodstock inventory management. Toward this end, I utilized a genotyping-by-sequencing (GBS) approach for bulk SNP sequencing and screening, followed by SNP validation and panel creation using MassARRAY technology [12-14].

2. Methods and Materials

2.1 Genotyping-by-Sequencing Sample Preparation and Sequencing

Fin clip samples were collected from 166 individuals representing one long-term domesticated population (nine generations of mass selection) housed at SNARC (initially sourced from Lake Erie and the Tennessee River) and five wild founder stocks sourced in 2015–2016 from the following locations and currently under initial evaluation at SNARC: Little Missouri River, AR (LMO), Ouachita River, AR (OUA), Tallapoosa River, AL (TAL), Neches River, TX (NTX), Nueces River, TX (STX). Sample numbers and GPS coordinates are given in Table 1. An additional 96 white bass individuals from the same geographic locations were sampled and used for genetic assignment analyses.

Genomic DNA was extracted from fin clips using the DNeasy Blood & Tissue kit (Qiagen), according to the manufacturer's protocol. The DNA quality was assessed by running 100 ng of each DNA sample on 1% agarose gels. DNA samples were sent to the University of Minnesota Genomics Center for GBS library construction and sequencing. Briefly, 100 ng of DNA was digested with 10 units *Bam*HI-HF (New England Biolabs; NEB) and incubated at 37 ^{o}C for 2 h. Samples were then ligated with 200 units of T4 ligase (NEB) and phased adaptors with GATC overhangs at 22 °C for 2 h and heat-killed. The ligated samples were purified with SPRI beads and then amplified for 18 cycles with 2×NEB HiFi Master Mix to add barcodes. Libraries of 166 white bass samples were purified, quantified, pooled and size selected for the 300- to 744-bp library region and diluted to 1.7 pm for sequencing. The pooled libraries were loaded across four lanes of a 150-bp single-read sequencing run on the NextSeq 550 instrument, with 47.1 million reads generated from sequencing.

Table 1 Geographic coordinates and number of individuals of six white bass populations

Group	Latitude	Longitude	Location	*n*
DOM	34.476	-91.418	Domesticated; SNARC	36
LMO	34.239	-93.658	Little Missouri River, AR	30
NTX	32.315	-95.466	Neches River, TX	25
OUA	34.626	-93.658	Ouachita River, AR	28
STX	28.430	-98.182	Nueces River, TX	13
TAL	32.956	-85.692	Tallapoosa River, AL	34

2.2 SNP Marker Development and Genotyping

To perform reference-based SNP calling and facilitate subsequent panel design, I assembled the first rough draft genome for white bass. An *M. chrysops* DNA sample was sequenced in a single lane of a HiSeq 2500 High Output 125-bp paired-end run. I assembled 265 million Illumina reads (~54-fold coverage) into a 621 Mb (N50 = 53.5 kb) draft genome using abyss v2.0.2 [15] with default parameter settings. Genome-wide SNPs were identified using tassel 5.0 GBS pipeline V2 [16] with default parameter settings. Vcftools [17] and genepop [18] were used for stringent filtering of SNPs based on the following criteria: (i) only biallelic SNPs were kept, (ii) SNPs had to be called in at least 80% of individuals, (iii) SNPs with minor allele frequency lower than 0.05 were removed, (iv) SNPs with observed heterozygosity larger than 0.6 were removed, (v) SNPs deviating from Hardy-Weinberg equilibrium in more than four populations were removed (P <

0.05) and (vi) SNPs in linkage disequilibrium were pruned after Bonferroni correction (applied during assay design).

A MassARRAY System (Agena Bioscience, San Diego, CA) was employed to validate and evaluate the performance of selected SNPs across white bass populations that included both GBS (n = 71) and additional individuals (n = 96) from the same geographic locations. SNP assays were designed using the MassARRAY Assay Design Software with the goal of maximizing multiplexing of 40 SNPs per well. During genotyping, amplification and extension reactions were performed using 10 ng of DNA per sample and utilizing the iPLEX Gold Reagent Kit according to the manufacturer's protocols. SNP genotypes were called using the MassARRAY Typer 4 analysis software. This software uses a three-parameter (mass, peak height and signal-to-noise ratio) model to calculate the significance of each genotype. A final genotype was called and assigned a particular name (e.g. conservative, moderate, aggressive, user call) based on probability. Noncalls were also noted (e.g. low probability, bad spectrum).

2.3 Population Genetic Analysis Using GBS Data

Population diversity and differentiation were evaluated by computing the observed (H_o) and expected heterozygosity (H_e), percentage of polymorphic loci (P_o) and pairwise fixation indices (F_{st}) using GenALEx 6.5 [19]. Phylogenetic relationships between white bass populations were analyzed based on pairwise F_{st} values using the UPGMA method in MEGA7.0 [20]. Population structure of white bass individuals was inferred using a systematic Bayesian clustering approach implemented in STRUCTURE v2.3.4 [21]. Data analysis was performed using the admixture and correlated allele frequencies models with a burn-in of 50 000 iterations followed by 500 000 repetitions of Markov chain Monte Carlo simulation. I used different numbers of assumed population genetic clusters (K = 1-10) to determine the best-supported *K* value using the webserver STRUCTURE HARVESTER [22], repeated 10 times for each K value. The structure results were summarized using CLUMPAK [23].

2.4 Population Assignment and Structure Analysis Using SNP Panels

The development of multiplexed SNP panels represents a valuable tool for white bass genetic variation analyses. I sought to ensure that the 57 selected SNPs on the current panels accurately approximated the STRUCTURE results obtained from the 426-SNP set (using 71 individuals genotyped by both GBS and MassARRAY) and, furthermore, that additional samples could be accurately assigned to their source populations.

To evaluate the genetic assignment performance of the SNP panels, GENECLASS v2 [24] was used to assign 96 additional fish to populations. SNP genotypes of six GBS populations were used as the reference for genetic assignments. A Bayesian classification method [25] implemented in geneclass2 was used to compute assignment likelihood scores of each individual fish for each reference population. White bass individuals were assigned to potential origin populations with the highest likelihood scores and a threshold score of 0.05 [26].

3. Results

3.1 SNP Marker Discovery and Genotyping

During GBS analysis, TASSEL generated 13,872 unfiltered SNP loci represented across six populations. After stringent filtering (criteria i-v), a total of 426 SNPs remained for subsequent multiplex panel design (Fig. 1). Following marker development, the MassARRAY System was employed to validate and evaluate the performance of selected SNPs using repeated GBS samples (n = 71) and additional white bass (n = 96) that sampled from the same geographic locations. Using MassARRAY Assay Design Software, I designed and ordered two multiplex assays, with 40 and 34 SNPs, respectively (Table 2). Of the 74 SNPs, 57 performed well based on genotype cluster distribution and concordance. Failing SNPs were potentially due to paralogous loci unidentified during primer design [27]. Comparison of the SNP genotypes derived from GBS and MassARRAY using 71 repeated individuals revealed a high concordance (97.84%), indicating high reliability and reproducibility in the selected markers.

Table 2 Primer sequence information for the two MassARRAY multiplex panels in white bass.

SNP_ID	SNP Alleles		Primer Sequences
S1280496_21508	C/T	PCR1:	ACGTTGGATGTTTCCACTGGATCCCCTTTC
		PCR2:	ACGTTGGATGCCGTATGTTGTAGGTGTTCC
		EXT:	ACCTCTCTTCTCCGT
S1270917_10924*	T/G	PCR1:	ACGTTGGATGAAGCACTACAGGTGATCAGC
		PCR2:	ACGTTGGATGCCCTTTCTTTTACCTGTGCG
		EXT:	TCCCAGGTGAGAAGG
S1249339_28049	C/T	PCR1:	ACGTTGGATGGCCTGATCTAAACCACTACG
		PCR2:	ACGTTGGATGGGGTCATGGGAGTAGTAATC
		EXT:	CACCGAGGATCCATTC
S1281292_84146	G/A	PCR1:	ACGTTGGATGTTATGTCCATCTGTGTCCTC
		PCR2:	ACGTTGGATGGATGTTTTTGGTACCCGCTG
		EXT:	gGTGTCCTCACAGCTC
S1273338_427*	G/T	PCR1:	ACGTTGGATGTCTTCTGCCTCTTGCTGGG
		PCR2:	ACGTTGGATGGGTCAAACTGGAGAGCGTC
		EXT:	ctTCTCCCCCAGCTTCG
S1277146_1383	C/A	PCR1:	ACGTTGGATGTGCAGCAGGATATCATTCCC
		PCR2:	ACGTTGGATGCAAAACCCTCTCACCCAAAC
		EXT:	ACTTGTTTGTTTCAGGC
S1280413_4483*	A/C	PCR1:	ACGTTGGATGATCCACTGCCTGACTTCAGC
		PCR2:	ACGTTGGATGCAGTGTCGGACGTAGATCTG
		EXT:	ACGTAGATCTGGTTTCG
S818106_183	G/A	PCR1:	ACGTTGGATGTGTAGTGCAGTCCCTTGTTG
		PCR2:	ACGTTGGATGGGACATTACAGTCCTCTTGC
		EXT:	CCCCTGCAGTTTTCCTAC
S1257462_11573	G/A	PCR1:	ACGTTGGATGCCACATGAGTGCAAGTAAAC
		PCR2:	ACGTTGGATGTCTGGATCCAGGTCATAGAG

SNP_ID	SNP Alleles		Primer Sequences
		EXT:	gtTGTGAGGTGCTTCAGT
S1277230_305	A/T	PCR1:	ACGTTGGATGGCAGTGGTAATAAAGAGAGG
		PCR2:	ACGTTGGATGTCTGGGACAAACACTGACTC
		EXT:	ATAAAGAGAGGTGGAGAC
S1275888_24859	A/G	PCR1:	ACGTTGGATGATTAGGGTCAGTGCTATCCG
		PCR2:	ACGTTGGATGTGAATGTGACAGCAGGGTTC
		EXT:	TGCTATCCGATATCGATTG
S1276194_62711	A/C	PCR1:	ACGTTGGATGACGCCAAGGATCAGAGTGAG
		PCR2:	ACGTTGGATGACCTCTGTCCTGTTAACTCG
		EXT:	aGCTAGCTTAGATGTTGCC
S1242368_7131	T/C	PCR1:	ACGTTGGATGCCTTTTGTTCTCTTCTAGGC
		PCR2:	ACGTTGGATGCAGTGGATCTCAGTCTCAAG
		EXT:	tattGATCCCACTCCACCCT
S1275465_16542*	T/G	PCR1:	ACGTTGGATGTACAGCGAGTGTGAGAAGAG
		PCR2:	ACGTTGGATGTGAGGATCCTGGGAAGAAAC
		EXT:	gGTGAGAAGAGGAGAAAGG
S1280383_71963	C/T	PCR1:	ACGTTGGATGACCAGCCTGTTTGTCTGTTG
		PCR2:	ACGTTGGATGACTGACTCCCTCTGTTCTTC
		EXT:	tgtaTGTTGGGGAAAACAGT
S1277609_119571	G/A	PCR1:	ACGTTGGATGAGAGAGCTGCTTGTGTAATC
		PCR2:	ACGTTGGATGGTTTAACATGGATCCCAAGC
		EXT:	cAATCAGAGTAGAGTGAAGG
S300698_211	C/T	PCR1:	ACGTTGGATGGTCTCCAGAGGCTGTAAATG
		PCR2:	ACGTTGGATGGCACTATAATGTCTTTACAGG
		EXT:	GGCTGTAAATGCCACAACAAC
S1275066_18277	T/A	PCR1:	ACGTTGGATGAAACGCTCTCAAGGTAGAGG
		PCR2:	ACGTTGGATGGGGCTCCAAACCAATATCTG
		EXT:	cGCTCTGCATTAGTGGTTAAA
S1279357_88278*	T/G	PCR1:	ACGTTGGATGGTTACGACGAGGCTTTGGAG
		PCR2:	ACGTTGGATGTAAAGCCCCCCTTGAAGAAG
		EXT:	ctttTTTTTTGTTGCTACCCCG
S1280341_56036	T/C	PCR1:	ACGTTGGATGTAATGATGTCAGTCCGTCCG
		PCR2:	ACGTTGGATGACATGAAAGACGTGGATCCG
		EXT:	tccggCCCAATGTCCACAGCGG
S1279254_7787	G/A	PCR1:	ACGTTGGATGAATGACCAGGTCTCTGCTTG
		PCR2:	ACGTTGGATGGAGATTAGACCACATGCCTG
		EXT:	TGGATCTGGATTAAATGTAAAG
S1280060_32824	G/A	PCR1:	ACGTTGGATGGAGTACTGTCCTTCATCAGC
		PCR2:	ACGTTGGATGATCCCTCATGTCCACAGAGC
		EXT:	ttcaTTGTCTTATCCATGCCACA
S1280755_8475	C/T	PCR1:	ACGTTGGATGTACAAGTGCAGACAATCCCC
		PCR2:	ACGTTGGATGAGTTCCAGCAAAGGCCACAG
		EXT:	agccAAGGCCACAGTGGATACAC
S1279169_35659*	T/G	PCR1:	ACGTTGGATGTCCGCCTCCAGGTATGGAAA

SNP_ID	SNP Alleles		Primer Sequences
		PCR2:	ACGTTGGATGGGTGTGATTTTCAGAACCGC
		EXT:	cctaACACCGGGATGGAGAAAGG
S1272940_825	T/C	PCR1:	ACGTTGGATGTCCAAGGTACAAGCAAAGTG
		PCR2:	ACGTTGGATGCATTAGGTTCTCCTACCCAC
		EXT:	ACTTGTATTACTACAAACTCTCAT
S1280807_77073	G/A	PCR1:	ACGTTGGATGGAGACCCTAGGATCCTATTG
		PCR2:	ACGTTGGATGACTGTGTCAATCAGTGCACC
		EXT:	TGCACCAATTATTGCTTCATCAAA
S1259785_2756	T/A	PCR1:	ACGTTGGATGAAAGCGGACCAGAGACAATG
		PCR2:	ACGTTGGATGGTCTGTAGTGGAGCAACGTG
		EXT:	CCAGAGACAATGATGTTATTATTT
S1278860_32228	A/T	PCR1:	ACGTTGGATGTCCCTTAATCGTCACTCTGC
		PCR2:	ACGTTGGATGATGCTGTTGTGTGAGCTAAG
		EXT:	ggaggCTGCCCAGAGTGCTCATAT
S1281175_89726	G/A	PCR1:	ACGTTGGATGCAAAATATGTACCGAGTAGG
		PCR2:	ACGTTGGATGCTGGATCCACTTGTAGACTG
		EXT:	ttttTCTTGTCTATCTGCTATCTGC
S1279936_99379	T/C	PCR1:	ACGTTGGATGAGAACTGCCATCACTGTGAG
		PCR2:	ACGTTGGATGCCAGGCCTACAGTGCTTATC
		EXT:	ttttCTACAGTGCTTATCATTCATC
S1258188_29724	G/C	PCR1:	ACGTTGGATGGTGAACTCCTCAGGATCCAC
		PCR2:	ACGTTGGATGAGCCTCTCACTTATTTCTCC
		EXT:	tCCTCTCACTTATTTCTCCACTTTTA
S1275674_1935	T/C	PCR1:	ACGTTGGATGGCTCAAAGTAATGCTTGAAC
		PCR2:	ACGTTGGATGTTGTGCAACCACCATCACAG
		EXT:	cATGCTTGAACTTAGCTTTTTTTTAT
S1278387_14770	G/A	PCR1:	ACGTTGGATGAAAGGATCCAAGTGAATGGG
		PCR2:	ACGTTGGATGCTTTCACTTCACTCAATTCC
		EXT:	gcatGGTAAACTCTGCATTTCCATAT
S1280073_41469*	T/G	PCR1:	ACGTTGGATGAGATCCTCACGTGGGTGCAG
		PCR2:	ACGTTGGATGGGATCCTAGATGATCCACTG
		EXT:	acggtGGGCCAACACCACCCCTCCTCC
S1277323_17323	G/A	PCR1:	ACGTTGGATGGGCCAAAAGCAATCACCCTT
		PCR2:	ACGTTGGATGTTGTTACAGTGCAGGATCCC
		EXT:	ttgttCAACATGGTACTAGGATCTCAT
S1276005_16740	G/T	PCR1:	ACGTTGGATGCCTCCCTTGTGGTATGCTTT
		PCR2:	ACGTTGGATGAGGACGCAGCAGTTGAAATC
		EXT:	GTTAAAGTGAATTTGAAAAAATCTATC
S1278634_131267	A/T	PCR1:	ACGTTGGATGATCCATTCCACTCCACATCC
		PCR2:	ACGTTGGATGCTCCATGGAAGGTCAATATG
		EXT:	ggggcCTCCACATCCCTGCCACGCTTAT
S1280142_15747	G/A	PCR1:	ACGTTGGATGAGCTCTTCTCCTTCATGCTG
		PCR2:	ACGTTGGATGGAAAGGATCCCCATCAAGTC
		EXT:	ggataCCCATCAAGTCCTCAAATGACAC

SNP_ID	SNP Alleles		Primer Sequences
S1233454_13342	C/T	PCR1:	ACGTTGGATGGTCAGTGCTGTACACAGAAA
		PCR2:	ACGTTGGATGTATACAGAAGTCTCCTGTGC
		EXT:	ggGCTGTACACAGAAAATTTAGTTCTAA
S1279833_50478	A/T	PCR1:	ACGTTGGATGGCTGTTAGCAGCAGAGTTTC
		PCR2:	ACGTTGGATGCGGATCCGTCTTAATGTACC
		EXT:	gaagtTTGTTGGTCTGATTCTAAAGTTA
S1280056_3938*	T/G	PCR1:	ACGTTGGATGCTTAAGTCATGGTACTCGGC
		PCR2:	ACGTTGGATGTGTGTCTCAAACAGGCATCC
		EXT:	AGGCATCCGCTCCCC
S1277172_7766*	T/G	PCR1:	ACGTTGGATGTCATTCCCTCTCTCCCGAAA
		PCR2:	ACGTTGGATGATCAAAGAGGATCTGCTCCC
		EXT:	CCCTGCTCTGTGGCG
S1273266_45557*	T/G	PCR1:	ACGTTGGATGTCAGAGTCGAACCCATGGTC
		PCR2:	ACGTTGGATGTCCTCATAATGTACCATGGC
		EXT:	ACCAGCTCCACAGGG
S1275251_163413*	T/G	PCR1:	ACGTTGGATGTGGTACCAACTTGTTGTTCG
		PCR2:	ACGTTGGATGATAGGGATGGATCCCCAGAC
		EXT:	GTTGTTCGAGCAGGG
S1275224_54013*	A/C	PCR1:	ACGTTGGATGTCCATCAATCACCTTGTCCC
		PCR2:	ACGTTGGATGTGGTTCTTGACCCAATGTCG
		EXT:	aCCAATGTCGGCTGGG
S1277609_94808	C/G	PCR1:	ACGTTGGATGTAAACCTGAGGGATCCACTG
		PCR2:	ACGTTGGATGGGACCAAAGACCACTATTGC
		EXT:	CGGTTTCATGCTGTTCA
S1232467_1411*	A/C	PCR1:	ACGTTGGATGGATGCAGCTCTATAGCATCC
		PCR2:	ACGTTGGATGTGGAGCGTGGAGTTTGATAG
		EXT:	TCATCAGAAATGGAGGG
S1275811_19329	A/C	PCR1:	ACGTTGGATGGATGCCTTTAGGAGAAGGAG
		PCR2:	ACGTTGGATGCTTCTATCCAGACTAATGGG
		EXT:	GGGGAGGTCCTTCCATCA
S1274706_11344	G/A	PCR1:	ACGTTGGATGGTCTAAGTTCCAGCATTCCC
		PCR2:	ACGTTGGATGTTTCGTCTGAACTCAGCTGG
		EXT:	CATTCCCATTCTTCTTTCA
S1266045_13214	C/T	PCR1:	ACGTTGGATGTGTCAATGATTCGTATGCTG
		PCR2:	ACGTTGGATGGCTTTCAAGAAGGGTTAAGG
		EXT:	GCTGTACCTTACATCACTG
S1275479_21941	T/A	PCR1:	ACGTTGGATGTTAGCCTACAGAAGGATCCC
		PCR2:	ACGTTGGATGGGGATGATGAATAGGGAATG
		EXT:	ACACATAAATGCATGCTCA
S205688_221	T/C	PCR1:	ACGTTGGATGGGCCGATATGATTAATTGGG
		PCR2:	ACGTTGGATGTGCTTTGATCCTCTGGTTGC
		EXT:	gtgGGCGACAGAAGAATCC
S1278094_5085	A/G	PCR1:	ACGTTGGATGCGTCTTGATTTTGAGGACCG
		PCR2:	ACGTTGGATGCATCTGCTAGCTCCTCACAC

SNP_ID	SNP Alleles		Primer Sequences
		EXT:	tgagCCTCACACCTCCTCAC
S1217564_353*	A/C	PCR1:	ACGTTGGATGGGACCAGGATCCAGTTTCAC
		PCR2:	ACGTTGGATGAGTCCTGCAGGGACGAAAC
		EXT:	ccAGAGGGAATGTTTCCTCA
S1243006_110831	A/G	PCR1:	ACGTTGGATGTCCTCCTGGCAGAAAATGAG
		PCR2:	ACGTTGGATGTCACGCTCCCTAAACATTCC
		EXT:	taCCTAAACATTCCCATCATG
S1275790_112412	A/G	PCR1:	ACGTTGGATGGGCAGAGGAGAACCATTATC
		PCR2:	ACGTTGGATGGCCACAAGATGGGAAAACTG
		EXT:	cttcTCCGAAGTCTGGACCCC
S1279102_33370	T/C	PCR1:	ACGTTGGATGTCTGTCACTGGATGGGTAAG
		PCR2:	ACGTTGGATGCATATAGCTGTAATGCAGAG
		EXT:	aacgGGATGGGTAAGACCACC
S1279762_44235	C/T	PCR1:	ACGTTGGATGTTTAAAGAGAAGTCGCCCCG
		PCR2:	ACGTTGGATGAGGTTTTGGCGCAGCTGAGT
		EXT:	gggtAATGCCGGGAGTCACAC
S909713_11652	A/C	PCR1:	ACGTTGGATGCAGAAAGGCGAAAATAACAC
		PCR2:	ACGTTGGATGCATGCCCCTTAATGCTAATC
		EXT:	ACACATTAATACCACATGCAAA
S1279222_9821*	G/T	PCR1:	ACGTTGGATGCTTGGTCCAAATTGTTGACT
		PCR2:	ACGTTGGATGACCAATAGAAGTCAGTGAGG
		EXT:	CCAAATTGTTGACTTTTATGAC
S1274835_22433	G/A	PCR1:	ACGTTGGATGAATCCCGTAACGCGTTTGAG
		PCR2:	ACGTTGGATGAAGTAGCTCTCAGACTTGGG
		EXT:	ttAAGCGCAGAGAGGAATATGG
S1267073_2296*	G/A	PCR1:	ACGTTGGATGGGGATCACCTTCACTTCATC
		PCR2:	ACGTTGGATGATGCATGAACTTCAGCACAC
		EXT:	agacACAGATGGCACTGGATCCT
S1278072_82224*	A/C	PCR1:	ACGTTGGATGCAGAGAGAAGACAGAAAGGG
		PCR2:	ACGTTGGATGTCAAGTACGGCCTGACAGAC
		EXT:	ACAGAAAGGGTTAAAAAAAAAAA
S1277249_1308	C/T	PCR1:	ACGTTGGATGAGGATCCCTTTTACCTCTGC
		PCR2:	ACGTTGGATGCCACTGATAATCTCACTGGT
		EXT:	ataCCTTTTACCTCTGCATTATTA
S1275013_25979	A/G	PCR1:	ACGTTGGATGTTTGGTGGATCCTGCTTGTC
		PCR2:	ACGTTGGATGTCCTCCAGTTAGAGCTAATG
		EXT:	cccgGTCAAATGTCTCACAACTAC
S1279968_5997	A/T	PCR1:	ACGTTGGATGTATGGCGACAATGCTTGACC
		PCR2:	ACGTTGGATGGGGAGGCGAACAGAAAAGAA
		EXT:	aagagTAAATCCTCAGCAGTCTCA
S1231416_12241	T/C	PCR1:	ACGTTGGATGGGATACTCCGCAGGTCTCT
		PCR2:	ACGTTGGATGTCCGTTTACGCGTCGTGCC
		EXT:	ccctaCGCAGGTCTCTGGAGACGCA
S1279107_30838	G/A	PCR1:	ACGTTGGATGGTTCACTAGAGGATCCCTGT

SNP_ID	SNP Alleles		Primer Sequences
		PCR2:	ACGTTGGATGGGTCATTAAAACAATATGAGG
		EXT:	GCTTTATGAAATCAATCCATTGTAT
S1275242_7423	G/C	PCR1:	ACGTTGGATGTCCACGCTGGAGTCTTTAAC
		PCR2:	ACGTTGGATGTATGGACCAGTCTGAGAACC
		EXT:	gagaTGGAGTCTTTAACGGACGTTG
S1276446_15293	G/A	PCR1:	ACGTTGGATGGTCATTGCTGGTTTGATCTG
		PCR2:	ACGTTGGATGACAAAAGCTTCAGAGACGGG
		EXT:	ccctcCTGCATGCCTGTCTGTCCTGG
S1273404_151032	C/G	PCR1:	ACGTTGGATGGCAGAAGTTTAGAGGCTCAG
		PCR2:	ACGTTGGATGTCTCTGTCTAAGACGTGCTC
		EXT:	ggttGCTGGCATTATTTTTGGATCCT
S1271713_11760	A/G	PCR1:	ACGTTGGATGAAAGACTGATCACCCTGCTG
		PCR2:	ACGTTGGATGATTGCTCCTTGTGGACTTGC
		EXT:	cccATTCTTTTCTGTCAAAATAAACTG
S1279005_17776	A/G	PCR1:	ACGTTGGATGGTAGTTCTCCCTGTGACAAG
		PCR2:	ACGTTGGATGGTCTAATGAGTGCTAGCAAC
		EXT:	GCAACTCAATATTACTATTACTAGTAC
S1274783_6967	C/T	PCR1:	ACGTTGGATGGAGAGCAGTCTGCAAAAAAC
		PCR2:	ACGTTGGATGCATGCTGCTGTGTAATGATG
		EXT:	ggGCAGTCTGCAAAAAACAAAAAAATC

* Excluded from SNP panels for genotyping

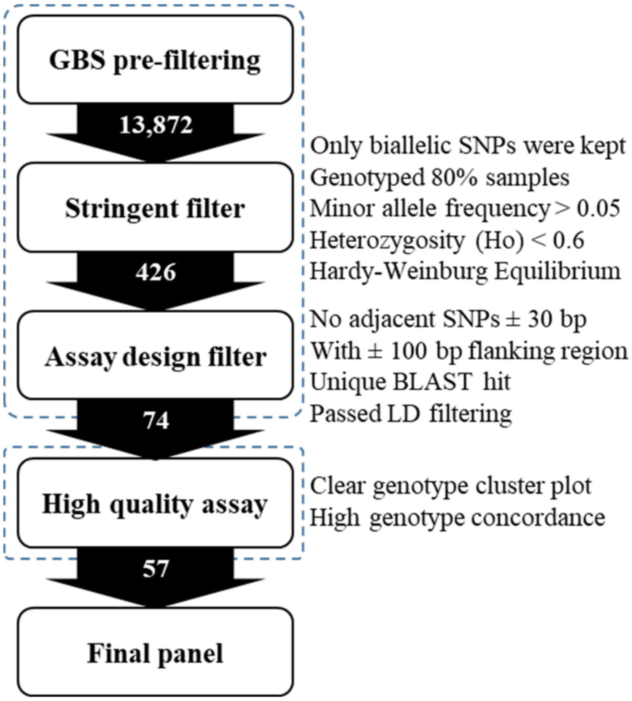

GBS pre-filtering
13,872
Stringent filter
Only biallelic SNPs were kept
Genotyped 80% samples
Minor allele frequency > 0.05
Heterozygosity (Ho) < 0.6
Hardy-Weinburg Equilibrium
426
Assay design filter
No adjacent SNPs ± 30 bp
With ± 100 bp flanking region
Unique BLAST hit
Passed LD filtering
74
High quality assay
Clear genotype cluster plot
High genotype concordance
57
Final panel

Fig.1 The workflow outlining the steps used in marker discovery and selection of SNP panels for population genetic analysis in white bass

3.2 Genetic Diversity and Population Structure in White Bass

Calculation of genetic diversity using 426 SNPs revealed similar marker polymorphism and heterozygosity rates among all examined populations (Table 3). Pairwise population F_{st} values ranged from 0.037 to 0.128, with a global F_{st} of 0.083 (Table 4). The minimum F_{st} of 0.037 was observed between LMO and NTX populations. The maximum pairwise F_{st} of 0.128 was detected between the domesticated (DOM) and OUA white bass populations.

Table 3 Genetic diversity parameters of six white bass populations based on 426 filtered SNP loci. P_o indicates percentage of polymorphic loci. H_o indicates average observed heterozygosity. H_e indicates average expected heterozygosity.

Group	Location	*n*	P_o %	H_o	H_e
DOM	Domesticated; SNARC	36	87.79	0.20	0.19
LMO	Little Missouri River, AR	30	95.07	0.20	0.20
NTX	Neches River, TX	25	91.08	0.20	0.19
OUA	Ouachita River, AR	28	85.45	0.19	0.19
STX	Nueces River, TX	13	80.99	0.19	0.19
TAL	Tallapoosa River, AL	34	92.49	0.20	0.20
Mean			88.81	0.20	0.19
SD			5.13	0.01	0.01

Table 4 Pairwise F_{st} values among six white bass populations based on 426 SNP loci.

	DOM	LMO	NTX	OUA	STX
DOM	–				
LMO	0.076	–			
NTX	0.118	0.037	–		
OUA	0.128	0.050	0.089	–	
STX	0.113	0.051	0.065	0.107	–
TAL	0.068	0.066	0.084	0.100	0.080

The phylogenetic analysis revealed two major genetic clusters of white bass populations: (i) DOM and TAL and (ii) populations from Arkansas and Texas (Fig. 2a). For all 166 GBS individuals from the six populations, STRUCTURE analysis with K = 2 received the strongest support (Fig. 2b). The same optimal K = 2 was obtained based on genotypes from 96 additional samples on the 57-SNP panel (Fig. 2c).

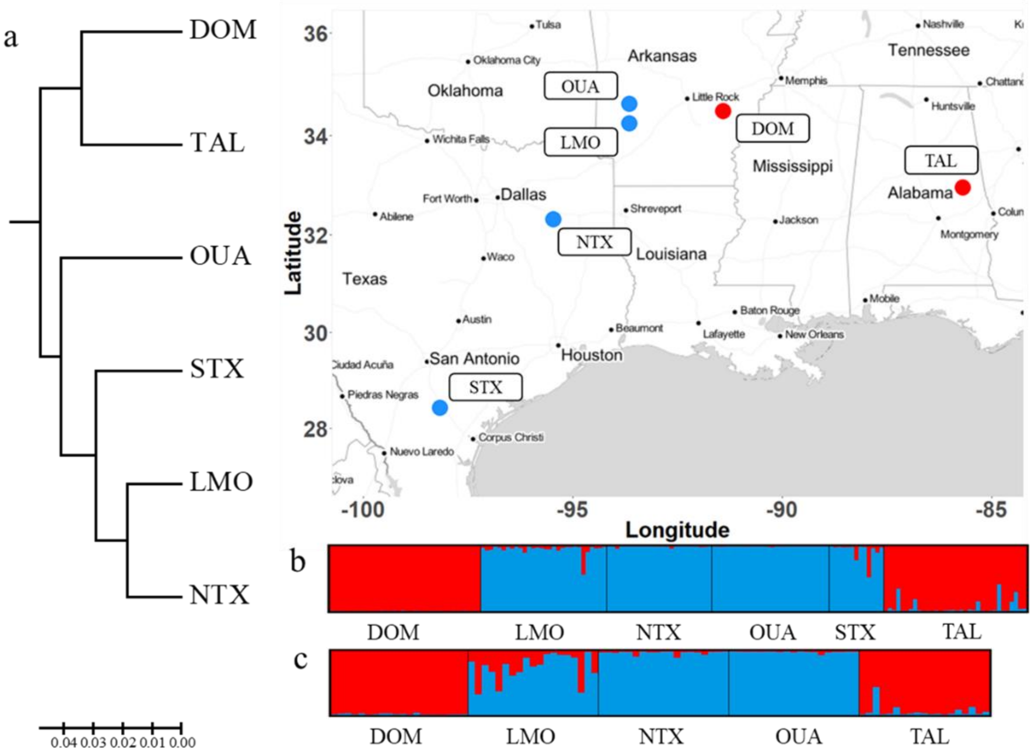


Fig. 2 (a) Evolutionary relationship of white bass populations based on pairwise population F_{st} values using UPGMA method. (b) STRUCTURE bar plot result of a total of 166 GBS individuals from six populations (K = 2). (c) STRUCTURE bar plot result of additional individuals from five of these populations (excluding STX due to lack of additional samples, K = 2) genotyped with 57 SNP panels.

3.3 SNP Panel for Population Assignment and Structure Analysis

The development of multiplexed SNP panels represents a valuable tool for white bass genetic variation analyses. I sought to ensure that the 57 selected SNPs on the current panels accurately approximated the STRUCTURE results obtained from the 426-SNP set (using 71 individuals genotyped by both GBS and MassARRAY) and, furthermore, that additional samples could be accurately assigned to their source populations. STRUCTURE q-values from genotypes of the 71 individuals were highly correlated between the 426-SNP GBS data and the 57-SNP MassARRAY data (r^2 = 0.958; data not shown).

To evaluate the genetic assignment performance of the SNP panels, I used GENECLASS v2 to assign 96 additional fish to populations. As shown in Table 5, All 96 fish were correctly assigned to their origin populations (100% accuracy) with an average assignment score of 99.78%, indicating that the 57 SNPs possess sufficient power for genetic assignment and discrimination among domesticated and founder stock white bass populations.

Table 5 Results of genetic assignment for additional 96 genotyped white bass individuals with known origin. GBS genotype data based on 426 SNPs and six populations were used as the reference. Numbers in bold represent the count of fish assigned to the correct origin. Additional samples were unavailable for the STX population.

Population	DOM	LMO	NTX	OUA	STX	TAL	Correctly	Average
DOM	**20**						100.00%	100.00%
LMO		**19**					100.00%	99.06%
NTX			**19**				100.00%	99.85%
OUA				**19**			100.00%	99.99%
TAL						**19**	100.00%	99.99%
Overall							100.00%	99.78%

4. Discussion

The results indicate that wild white bass sourced from Texas rivers (where they were introduced) share recent ancestry with wild fish from Arkansas rivers (within the species' native range), suggesting that Texas white bass populations were likely derived from lower Mississippi River drainages. On the other hand, DOM and TAL individuals may be clustered due to shared ancestry derived from the more isolated Tennessee River. Records indicating the source populations for initial stockings into NTX, STX and TAL localities appear to be non-existent. Future genetic analyses of wild white bass populations in Lake Erie, the Tennessee River and the upper Mississippi River may shed further light on the source of introduced white bass populations.

The SNP GBS dataset and validated SNP panels described here expand the genetic toolbox for white bass. Although most immediately aiding the white bass selective breeding program, these resources should also be of keen interest to state and federal natural resource agencies that wish to better understand and manage native and introduced populations of this important species.

References

1. Beck, B.H., et al., Hepatic transcriptomic and metabolic responses of hybrid striped bass (*Morone saxatilis*× *Morone chrysops*) to acute and chronic hypoxic insult. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics, 2016. **18**: p. 1-9.
2. White, M.M., Genetic variation in white bass. Transactions of the American Fisheries Society, 2000. **129**(3): p. 879-885.
3. Fuller, S.A., et al., Heritability of growth traits and correlation with hepatic gene expression among hybrid striped bass exhibiting extremes in performance. Cogent Biology, 2018. **4**(1): p. 1453319.
4. Garber, A.F. and C.V. Sullivan, Selective breeding for the hybrid striped bass (*Morone chrysops*, Rafinesque× *M. saxatilis*, Walbaum) industry: status and perspectives. Aquaculture Research, 2006. **37**(4): p. 319-338.
5. Hallerman, E.M., Toward Coordination and Funding of Long‐Term Genetic Improvement Programs for Striped and Hybrid Bass Morone sp. Journal of the World Aquaculture Society, 1994. **25**(3): p. 360-365.
6. Kenter, L.W., et al., Strain evaluation of striped bass (*Morone saxatilis*) cultured at different salinities. Aquaculture, 2018. **492**: p. 215-225.
7. Anderson, A.P., M.R. Denson, and T.L. Darden, Genetic structure of striped bass in the southeastern United States and effects from stock enhancement. North American Journal of Fisheries Management, 2014. **34**(3): p. 653-667.

8. Couch, C., et al., Isolation and characterization of 149 novel microsatellite DNA markers for striped bass, *Morone saxatilis*, and cross‐species amplification in white bass, *Morone chrysops*, and their hybrid. Molecular Ecology Notes, 2006. **6**(3): p. 667-669.

9. Li, C., et al., Transcriptome annotation and marker discovery in white bass (*Morone chrysops*) and striped bass (*Morone saxatilis*). Animal genetics, 2014. **45**(6): p. 885-887.

10. Fuller, S.A., B.D. Farmer, and B.H. Beck, White bass *Morone chrysops* is less susceptible than its hybrid to experimental infection with *Flavobacterium columnare.* Diseases of Aquatic Organisms, 2014. **109**(1): p. 15-22.

11. Fuller, S.A. and M.M. McEntire, Variation in body weight and total length among families of white bass, *Morone chrysops*, fry after communal rearing. Journal of Applied Aquaculture, 2011. **23**(3): p. 250-255.

12. Zhao, H., et al., SNP marker panels for parentage assignment and traceability in the Florida bass (*Micropterus floridanus*). Aquaculture, 2018. **485**: p. 30-38.

13. Thongda, W., et al., Development of SNP panels as a new tool to assess the genetic diversity, population structure, and parentage analysis of the eastern oyster (*Crassostrea virginica*). Marine Biotechnology, 2018. **20**(3): p. 385-395.

14. Li, C., et al., Discovery and validation of gene‐linked diagnostic SNP markers for assessing hybridization between Largemouth bass (*Micropterus salmoides*) and Florida bass (*M. floridanus*). Molecular Ecology Resources, 2015. **15**(2): p. 395-404.

15. Jackman, S.D., et al., ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Research, 2017. **27**(5): p. 768-777.

16. Glaubitz, J.C., et al., TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One, 2014. **9**(2): p. e90346.

17. Danecek, P., et al., The variant call format and VCFtools. Bioinformatics, 2011. **27**(15): p. 2156-2158.

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

19. Peakall, R. and P.E. Smouse, GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes, 2006. **6**(1): p. 288-295.

20. Kumar, S., G. Stecher, and K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution, 2016. **33**(7): p. 1870-1874.

21. Pritchard, J.K., M. Stephens, and P. Donnelly, Inference of population structure using multilocus genotype data. Genetics, 2000. **155**(2): p. 945-959.

22. Earl, D.A., STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 2012. **4**(2): p. 359-361.

23. Kopelman, N.M., et al., Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources, 2015. **15**(5): p. 1179-1191.

24. Piry, S., et al., GENECLASS2: a software for genetic assignment and first-generation migrant detection. Journal of Heredity, 2004. **95**(6): p. 536-539.

25. Rannala, B. and J.L. Mountain, Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences, 1997. **94**(17): p. 9197-9201.

26. Epps, C.W., et al., Using genetic tools to track desert bighorn sheep colonizations. The Journal of Wildlife Management, 2010. **74**(3): p. 522-531.

27. Andrews, K.R., et al., Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 2016. **17**(2): p. 81-92.

Chapter IV SNP discovery and panel development for genetic identification and hybridization analysis in walleye (*Sander vitreus*)

Abstract

Walleye (*Sander vitreus*) is a freshwater fish species inhabiting a wide range of habitats across North America. Owing to its economic importance and popularity as a sportfish, this species has been widely introduced beyond its indigenous range. Previous genetic studies have revealed a genetically distinct group of walleye in Mobile River Basin drainages, but little is known about its genetic structure and how this southern assemblage of populations relates to northern walleye populations. Another unresolved question is whether walleye in Mobile River Basin have interbred with northern walleye following introductions. In this study, I used genotyping-by-sequencing (GBS) data from 60 walleye individuals to infer genetic diversity and structure among northern and southern (Mobile River Basin) walleye populations. Genetic assessment with 2,782 GBS SNPs confirmed a unique genomic pattern in the southern walleye group (F_{ST} = 0.805 vs. northern Lake Erie walleye). I also found strong evidence for a historical declining population trend with reduced genetic diversity and effective population size in a southern walleye population spawning in Hatchet Creek, Alabama. Additionally, a SNP assay with 68 diagnostic markers was developed for rapid and accurate identification of genetic purity and classification of various (northern/southern) hybrid classes among walleye individuals. Utilizing this SNP assay, I genotyped an additional 545 walleye individuals across 23 populations and characterized broad-scale genetic structure, distinguishing three groups comprised of the Mobile River Basin, Great Lakes/upper Mississippi, and southern Ohio drainage populations. Using both GBS and SNP assay data, I demonstrated that a suite of 68 SNPs could collectively classify advanced-generation

hybrids and enable us to identify an anthropogenic hybrid zone resulting from the previous introduction of northern walleye into the Black Warrior River. The results highlight the need for further characterization and conservation management of southern walleye in the Mobile River Basin, with the 68-SNP assay currently being implemented in ongoing stream survey and captive breeding programs.

1. Introduction

Freshwater ecosystems are severely threatened by anthropogenic activities such as overexploitation, water pollution, destruction of habitats, and species invasion [1]. For exploited freshwater fishes of economic or recreational importance, restocking is commonly employed in order to increase population size and mitigate the risks of genetic collapse [2]. Effective restocking practices should take into account the existing population structure and genetic diversity of the species [3], however this is not always the case [4]. Population genetics is a powerful tool in conservation biology, as it can provide information to reduce risks from the loss of genetic diversity, inbreeding depression, and introgression with non-native individuals. Recent developments in genomic approaches, such as genotyping-by-sequencing (GBS), have facilitated the cost-effective detection of genome-wide single nucleotide polymorphisms (SNPs) [5]. Due to improved resolution for describing hybridization/introgression, adaptive genetic variation, and fine-scale demographic structure, SNPs are rapidly becoming the preferred markers for routine genetic identification and hybridization tests in a variety of aquatic species [6-10].

Walleye (*Sander vitreus)* is an ecologically important and economically valuable freshwater fish species in family Percidae [11]. Walleye inhabit a wide range of habitat conditions across North America, from the Mackenzie River in the Arctic to the U.S. Gulf Coast [11]. Its eastwest distribution is delineated by the eastern continental divide and the Rocky Mountains [12]. The distribution of walleye in southern latitudes is limited by the cold temperatures required for gonadal maturation [13]. Previous genetic surveys of walleye provided evidence for the existence of five genetically distinct lineages in North America, comprised of three northern stocks across the Northwest Lake Plains, Great Lakes watershed, and North Atlantic coastal drainages [14-16], and two unique lineages identified in the Mobile River Basin [17] and the eastern highland regions [18, 19]. A number of population genetic studies of walleye indicate that northern lineages were shaped by post-Pleistocene recolonization events [14, 16, 20], while populations from unglaciated Eastern Highlands regions evolved due to isolation in river drainages [19, 21]. Specifically, the southern populations in Alabama and Mississippi (Tombigbee River) were identified as a long-isolated historic assemblage of populations that diverged from northern walleye ~1.17 (± 0.31) Mya [19, 20]. Additional genetic investigations using allozymes [22, 23], mtDNA [17, 20, 22], and microsatellites [19, 24] confirmed their unique genetic pattern and suggested that a careful monitoring and management plan should be implemented for these southern populations due to potential introgression with introduced northern walleye [17]. Although the genetic divergence between Mobile River Basin walleye and other more northern walleye populations has been initially characterized, knowledge of genetic diversity, population structure and demographic history in southern populations is minimal. This is due, in large part, to the paucity of modern, verified genetic markers for the southern walleye.

Introgression has already been documented between introduced northern Great Lake stocks and southern walleye populations in the Ohio River drainage using mitochondrial DNA [21, 22, 25]. The same study by Billington and Maceina (1997) [26] also indicated that the integrity of southern walleye mtDNA haplotypes has been well preserved at Hatchet Creek (Coosa River drainage, Alabama). However, population monitoring suggested that this population was in decline and threatened by historical stocking of northern walleye. Ongoing walleye conservation efforts through restocking and captive breeding would benefit considerably from results of a comprehensive survey of existing genetic diversity and introgression among source populations, as well as the development of markers for rapidly characterizing genetic background.

Therefore, the primary goals of this study were to infer genetic diversity and population structure among northern and southern walleye populations using thousands of GBS-derived SNPs, and then develop a SNP panel that could rapidly and accurately distinguish between northern and southern walleye individuals and identify hybrids. To validate the SNP panel for rapid and accurate identification of various hybrid classes in walleye individuals, I focused on walleye from the Black Warrior River system in Alabama, as this system has a long history of non-native walleye stocking [26]. The secondary goal was to utilize the SNP panel for characterizing range-wide population structure and introgression in 23 walleye populations. Lastly, I compared historical demographic models using the GBS data to investigate whether a known pure southern walleye population shows genetic signatures of population decline.

2. Methods and Materials

2.1 Sample Collection and Genotyping-by-Sequencing

A total of 60 samples representing pure northern (Lake Erie, *n*=30, ERI), pure southern (Hatchet Creek, *n*=10, HAT) and hybrid (Blackwater Creek, *n*=20, BLA) walleye populations (Fig. 1) were collected for GBS library construction and sequencing. I also sampled an additional 545 walleye individuals across 23 northern and southern populations for extensive walleye population genetic structure analysis and hybrid classification (Fig. 1, Table 1). These additional individuals were collected to represent native walleye distribution at lacustrine and river sites, including the Great Lakes watershed (Lakes Erie, Michigan, Superior), Northwest Lake Plains (Mille Lacs Lake at the upper Mississippi River), Mobile River Basin (Coosa and Tombigbee River) and the eastern highlands regions (New River, Rockcastle River, and Big Sandy River). Meanwhile, sites with records of historical restocking (Black Warrior River and sites in the Tennessee River system) were also included in this study for further hybridization and introgression analyses. Fin clips of individuals were collected and stored in 95% ethanol for subsequent DNA extraction.

Table 1 Sample information for walleye populations used for GBS and MassARRAY data analyses.

Major	River	Population	Abbrev.	GBS	MassARRAY
Alabama	Coosa River	Hatchet Creek	HAT	10	46
		White Plains	WHI	0	11
	Tombigbee River	Tombigbee River	TOM	0	104
	Black Warrior River	Mulberry Fork	MUL	0	11
		Blackwater Creek	BLA	20	44
		North River	NOR	0	3
Tennessee	Little Tennessee River	Nantahala Lake	NAN	0	28
		Lake Fontana	FON	0	23
	Tennessee River	Normandy Lake	NOD	0	20
		Chickamauga Lake	CHI	0	20
		Watts Bar Lake	WAT	0	5
		Douglas Lake	DOU	0	5
		Cherokee Lake	CHE	0	5
		Norris Lake	NOS	0	5
		Fort Patrick Henry	PAT	0	5
Mississippi	Upper Mississippi	Mille Lacs Lake	MIL	0	39
Ohio	New River	Fosters Falls	FOS	0	50
	Rockcastle River	Rockcastle River	ROC	0	19
	Big Sandy River	Levisa Fork	LEV	0	8
Major	River	Population	Abbrev.	GBS	MassARRAY
Great Lakes	Lake Erie	Huron River	HUR	0	18
		Lake Erie	ERI	30	49
	Lake Michigan	Muskegon River	MUS	0	73
	Lake Superior	Thunder Bay	THU	0	14
Total				60	605

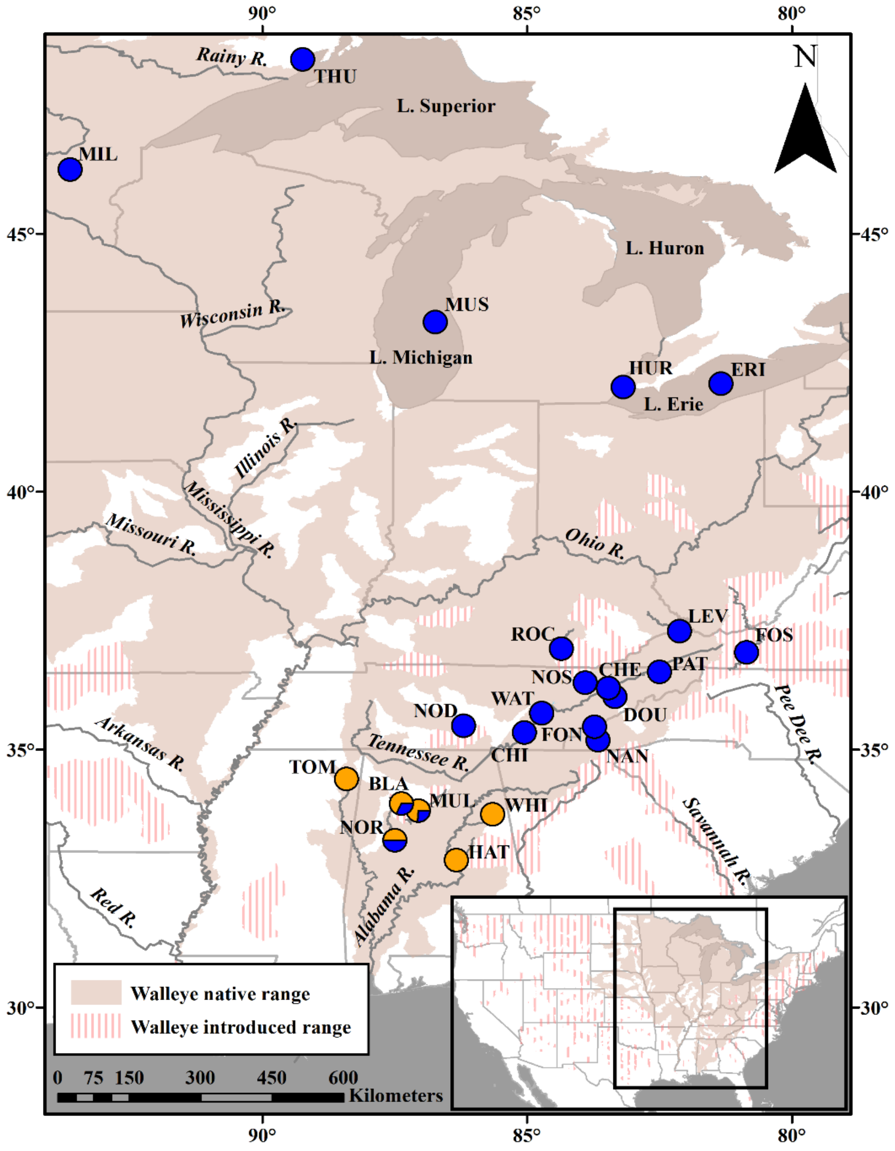


Fig. 1 Sampling locations of walleye. Populations are labeled as in Table 1 and represented by a pie graph showing the estimated southern and northern walleye genomic composition for each location based on STRUCTURE results with the 68-SNP panel (K=2).

Genomic DNA from all samples was extracted from fin clips using the DNeasy Blood & Tissue kit (Qiagen) according to the manufacturer's protocol. DNA quality was assessed by running 100 ng of each DNA sample on 1% agarose gels. DNA concentration was determined using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen). DNA samples were sent to the

University of Minnesota Genomics Center for double-digest GBS library construction and sequencing. Briefly, 100 ng of DNA was digested with 10 units of a combination of *Bam*HI and *Nsi*I enzymes (New England Biolabs; NEB) and incubated at 37 °C for 2 h. Following digestion, samples were then ligated with 200 units of T4 ligase (NEB) and phased adaptors at 22 °C for 2 h to inactivate the T4 ligase. The ligated samples were then amplified for 18 cycles with 2× NEB *Taq* Master Mix along with sample-specific barcodes. Libraries of walleye samples were purified, quantified, pooled and size selected for 300- to 744-bp fragments and diluted to 1.7 pm for sequencing. The pooled libraries were loaded across four lanes of 150-bp single-read sequencing on an Illumina NextSeq 550.

2.2 Genome Assembly and SNP Marker Discovery

To perform reference-based SNP calling, I assembled a rough draft genome for walleye. One DNA sample from Blackwater Creek walleye was selected for whole-genome sequencing and sent to the University of Minnesota Genomics Center for library construction and sequencing. During library creation, 100 ng of DNA was fragmented to target a 350-bp insert length using Covaris ultrasonic shearing. The sheared DNA was then end-repaired and subjected to a bead-based size selection. After adaptor and index ligation, the library was amplified using 8 cycles of PCR. The amplified library was sequenced across 1.5 lanes of a HiSeq 2500 125-bp paired-end run. A total of 302 million Illumina reads were generated from library sequencing and assembled into a 783 Mb (N50 = 4.13 kb) draft genome using MaSuRCA v3.2.4 [27]. I followed the default parameter settings for genome assembly, except for library insert length (342) and standard deviation (76). These two parameters were estimated using the Burrows-Wheeler Aligner (BWA) v0.7.17 [28].

Genome-wide SNPs were called using STACKS v2.4 [29], with minor changes in parameter settings. GBS reads were cleaned and de-multiplexed using the *process_radtags* program in STACKS. For reference-based SNP calling, I first mapped the de-multiplexed reads to the assembled walleye genome using BWA v0.7.17 [28]. The mapped reads were then sorted using the *sort* function in SAMTOOLS v1.6 [30]. The mapped and sorted reads were used to call SNPs with the *ref_map.pl* and *populations* pipelines in STACKS. I generated two SNP datasets in STACKS, one including all GBS samples (*n*=60) and the other with Hatchet Creek samples only (*n* =10). For population genetic analyses, I used the SNP dataset containing all samples, while the demography analysis was performed with SNPs from Hatchet Creek only. For the dataset including all GBS samples, I initially used VCFtools [31] to filter loci with minimum minor allele frequency (--maf) set to 0.05, minimum minor allele count (--mac) set to 10, and minimum locus coverage (--max-missing) set to 0.1. To ensure that SNPs were informative and reliable for downstream population genetic analyses and marker validation, VCFtools and SNPRelate [32] were used for stringent filtration of SNPs based on the following criteria: 1) only SNPs called in 100% of individuals; 2) SNPs with observed heterozygosity larger than 0.6 were removed to avoid paralogous loci in the dataset [33]; 3) SNPs deviating from Hardy-Weinberg equilibrium (HWE, p-value < 0.01) in more than one population were removed; 4) SNP pairs that showed linkage disequilibrium (LD) with $r^2 > 0.2$ were pruned (individual SNPs with higher genotype coverage were kept). For the demography dataset, I applied LD filtering at $r^2 = 0.2$ and kept only SNPs with no missing data.

2.3 Diagnostic Marker Development and Validation

In order to develop SNP assays for rapid and accurate identification of walleye lineages and various hybrid classes, I identified diagnostic SNPs with fixed-allelic differences between representative northern (Lake Erie) and southern (Hatchet Creek) populations (e.g., homozygous ‘A’ in pure southern individuals, homozygous ‘T’ in pure northern individuals, and polymorphic in Black Warrior River fish). I used GenAlex v6.5 [34] to identify putative neutral loci based on the distribution of SNP F_{ST} values (after exclusion of fixed markers). I ordered all SNPs based on locus-specific F_{ST} (from lowest to highest, across all population pairs) and created three data subsets that fell in different quartiles of the F_{ST} distribution: low global F_{ST} SNPs (below 25th percentile of the F_{ST} distribution corresponding to F_{ST} = 0.001-0.112), intermediate global F_{ST} dataset (between 25th and 75th percentile of F_{ST} distribution, F_{ST} = 0.112-0.378), and high global F_{ST} (75%-100% percentile, F_{ST}=0.378-0.822). An additional outlier scan was performed using the same dataset to identify SNPs showing evidence of divergent or balancing selection. The outlier test was conducted using BAYESCAN v2.1 with default iteration and burn-in settings, and prior odds set to either 1 or 10 [35]. Here the prior odds of 10 corresponds to a prior belief that the neutral model is 10 times more likely than the model of selection, while 1 represents the equal prior probability for both models. SNPs with a false discovery rate (FDR) <10% were considered as putatively under selection.

A MassARRAY System (Agena Bioscience, San Diego, CA) was used to validate a subset of diagnostic SNPs identified and genotyped by GBS and to genotype an additional 545 walleye individuals from 23 populations. Using MassARRAY ASSAY DESIGN software and following the protocol described in Zhao et al. (2018) [36], I designed two multiplex assays with 40 SNPs per well. Amplification and extension reactions were performed using 10 ng of DNA per sample and the iPLEX Gold Reagent Kit (Agena Bioscience) according to the manufacturer's protocol. SNP genotypes were called using the MassARRAY Typer 4 analysis software. This software uses a three-parameter (mass, peak height and signal-to-noise ratio) model to estimate genotype probabilities. Considering that all types of molecular markers are prone to genotyping errors [37], I used 59 individuals to test the concordance of SNP genotypes generated from MassARRAY and GBS. A total of 114 individuals (including 58 Black Warrior River samples) were used as technical replicates (genotyped twice by the MassARRAY system) to test the consistency of genotype calling in the MassARRAY system. Discordant genotypes due to missing data were excluded from this analysis.

2.4 Population Genetic Analyses

The Bayesian clustering algorithm-based program STRUCTURE v2.3.4 [38], was used to characterize population structure for both the GBS and MassARRAY datasets. The admixture model with correlated allele frequencies was applied with a burn-in of 20,000 iterations followed by 200,000 Markov Chain Monte Carlo (MCMC) repetitions. I used different numbers of assumed population genetic clusters (K=1-9 for GBS data, 1-25 for MassARRAY data) to determine the best-supported K value using the webserver CLUMPAK [39], repeated 10 times for each K. Population differentiation was estimated for all pairs of populations in the GBS and MassARRAY datasets using Hudson's estimator of F_{ST} [40] implemented in EIGENSOFT v7.2.1 [41]. Hudson's F_{ST} statistic is not sensitive to uneven population sizes and does not systematically overestimate F_{ST} [40, 42].

Population structure in the MassARRAY dataset was also visualized with a discriminant analysis of principal components (DAPC) implemented in the R package Adegenet [43]. The optimal number of principal components was determined by an alpha-score procedure with 20 repeated runs. For only the GBS dataset, population diversity indices for each population were evaluated by computing the observed heterozygosity (H_o), expected heterozygosity (H_e), and inbreeding coefficient (F_{is}) using Arlequin v3.5 [44]. Effective population size (N_e) for each population was estimated using the linkage disequilibrium method implemented in NeEstimator v2.1 [45].

2.5 Hybridization Analyses

One goal in this study was to investigate introgression and hybrid status in the Black Warrior River and evaluate the performance of the SNP panel for hybrid classification. For this purpose, I developed a novel framework that combined both custom and NEWHYBRIDS [46] methods for hybrid classification, with the assumption that no genotyping errors or contamination had occurred in genotyping data [6]. Briefly, before the hybrid assignment, a custom R script (available at https://github.com/hzz0024/walleye) was used to score three genotype ratios (homozygous AA and BB, heterozygous AB) for each examined fish. Individuals containing a mixture of homozygous loci for each parental group and heterozygous loci were manually assigned to later-generation backcross × backcross hybrids (e.g., F_xS with a majority southern walleye homozygous genotypes or F_xN with a majority northern walleye homozygous genotypes, Table 2), except for those consistent with F_2 proportions [6]. The Bayesian framework-based program, NEWHYBRIDS v1.1 beta, was then used to compute the posterior distribution of individual assignment into 12 different hybrid categories. The hybrid test was conducted using 300 fixed SNPs from the GBS dataset and 68 SNPs in MassARRAY assays, respectively. I trimmed the GBS datasets to 300 unlinked diagnostic SNPs because analyses failed to run with more markers due to an underflow issue [47]. Three independent analyses were conducted with different random subsets of 300 for assignment evaluation. I used HYBRIDLAB v1.0 [48] to simulate hybrids from northern and southern baseline populations in order to evaluate the power of the diagnostic SNPs in hybrid class discrimination. The program generated 15,000 random genotypes for each of the 12 hybrid classes (Table 2): parental northern walleye (ERI), parental southern walleye (HAT), F_1 hybrids, F_2 hybrids (F_1 × F_1), first-generation backcrosses (F_1 hybrids × either parental baseline), second-generation backcrosses (first-generation backcross × either parental baseline), third-generation backcrosses (second-generation backcross × either parental baseline) and fourth-generation backcrosses (third-generation backcross × either parental baseline). I did not include simulations of hybrids between backcross × backcross in the hybrid tests because of low assignment confidence among these classes [6]. I ran NEWHYBRIDS analyses using an initial 100,000 MCMC burn-in followed by 100,000 MCMC sweeps, with the 12 hybrid categories set as “Jeffreys-like priors” [46]. An individual was considered to be robustly classified if its assignment probability was > 0.5.

Table 2 Details of assignment criteria for southern, northern walleye and their hybrids identification. I used the same assignment criteria published in Lamer et al., 2015 [6]. The first twelve hybrid categories were used for NEWHYBRIDS analyses. The genotype probabilities within F_x categories (with both homozygous parental genotypes and heterozygous genotypes) were used for manual assignments.

NEWHYBRIDS category		Genotype probability		
		S (AA)	H (AB)	N (BB)
N	Pure northern walleye	0	0	1
S	Pure southern walleye	1	0	0
F1	First-generation hybrid	0	1.0	0
B_xN	First-generation backcross	0	0.5	0.5
B_xS	First-generation backcross	0.5	0.5	0
F2	Second-generation hybrid	0.25	0.50	0.25
B_x2N	Second-generation backcross	0	0.25	0.75

NEWHYBRIDS category		Genotype probability		
		S (AA)	H (AB)	N (BB)
B_x2S	Second-generation backcross	0.75	0.25	0
B_x3N	Third-generation backcross	0	0.125	0.875
B_x3S	Third-generation backcross	0.875	0.125	0
B_x4N	Fourth-generation backcross	0	0.0625	0.9375
B_x4S	Fourth-generation backcross	0.9375	0.0625	0
F_xS	Bx3S × Bx3S	0.8789	0.1172	0.0039
	Bx2S × Bx3S	0.8203	0.1719	0.0078
	Bx2S × Bx2S	0.7656	0.2188	0.0156
	BxS × Bx3S	0.7031	0.2813	0.0156
	BxS × Bx2S	0.6563	0.3125	0.0313
	BxS × BxS	0.5625	0.3750	0.0625
F_xN	Bx3N × Bx3N	0.0039	0.1172	0.8789
	Bx2N × Bx3N	0.0078	0.1719	0.8203
	Bx2N × Bx2N	0.0156	0.2188	0.7656
	BxN × Bx3N	0.0156	0.2813	0.7031
	BxN × Bx2N	0.0313	0.3125	0.6563
	BxN × BxN	0.0625	0.3750	0.5625
B_xBx	Bx3N × Bx3S	0.0586	0.8828	0.0586
	Bx2N × Bx3S	0.1172	0.8281	0.0547

2.6 Historical Demography Analyses of Hatchet Creek Walleye

Given the low genetic diversity and effective population size characteristic of southern walleye populations, I conducted model-based demographic analyses on Hatchet Creek samples using the composite likelihood approach implemented in FASTSIMCOAL v2.6.0.3 [49]. The goal here was to evaluate whether evidence exists for declines in population size and to gain a better understanding of the underlying temporal dynamic in this southern walleye population. The folded site frequency spectrum (SFS) was estimated from the STACKS dataset with only Hatchet Creek samples and used to compare four demographic models: constant population size, continuous population decline, instantaneous bottleneck, and a scenario of bottleneck followed by continuous decline (Fig. 2). These models were similar to those used in Chattopadhyay et al. (2019) [50], with minor modifications to the prior parameters (Fig. 2). In order to stabilize the estimated values, I conducted 50 independent runs for each model, and each run performed 100,000 simulations and

40 optimization cycles using a conditional maximization algorithm (ECM). The maximum-likelihood runs of each model were then compared and the Akaike information criterion (AIC), ΔAIC and the Akaike's weight were estimated to determine the best-fit demographical model [49]. Because there is no empirical estimate of genome-wide mutation rate for walleye, I applied different values of mutation rate (μ) for demographic model analyses. Three genome-wide mutation rates, human (2.5×10^{-8} per site/generation, [51]), cichlid (3.5×10^{-9} per site/generation; [52]) and Atlantic herring (2×10^{-9} per site/generation; [53]) were used to represent "high", "intermediate" and "low" mutation levels for model likelihood estimation and comparison. Ultimately, I chose the Atlantic herring mutation rate as it resulted in more similar estimated and observed likelihood values.

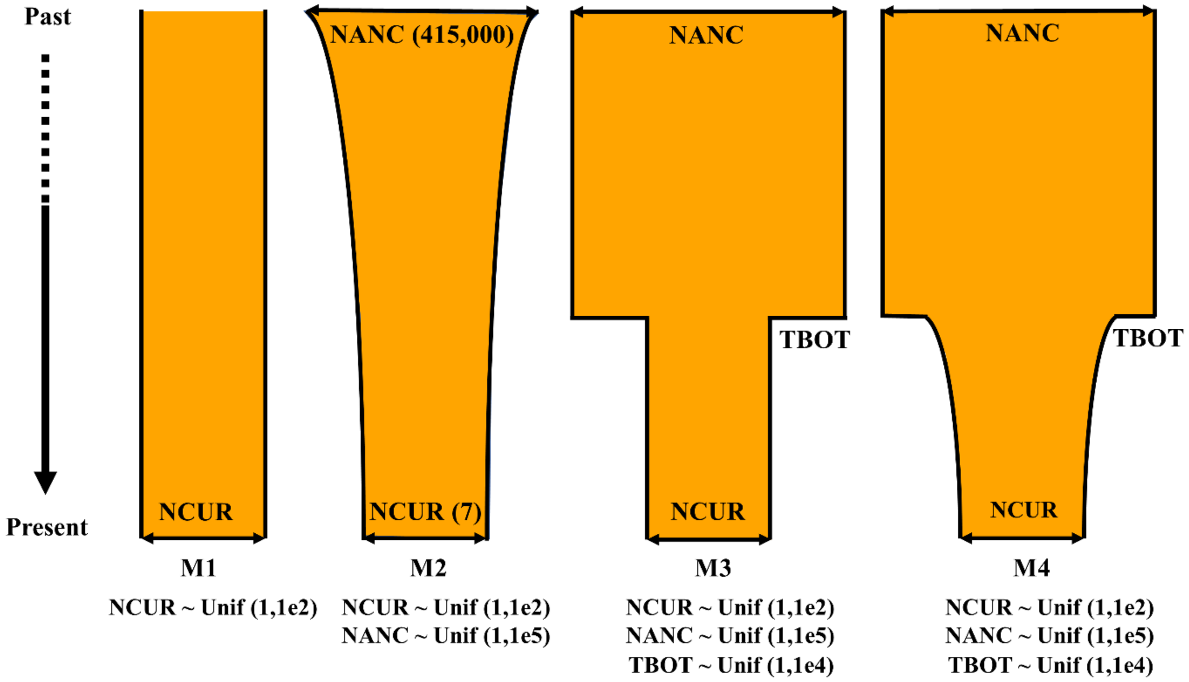


Fig. 2 Demographic models utilized to test declines in the southern walleye population using FASTSIMCOAL. Models from left to right: M1, a consistent model assuming no population size change over time; M2, a continuous decline model; M3, an instantaneous bottleneck model; and M4, a scenario of bottleneck followed by continuous decline. NCUR, the current population size; NANC, the ancestor population size or population size before bottleneck; TBOT, the time of

instantaneous population size change. The average estimated effective population sizes are shown at various stages in the best model (M2, in the bracket).

After model determination, I performed an additional 50 FASTSIMCOAL runs with a fixed ancestral N_e to obtain confidence limits for parameter estimates, choosing the parameter estimates from the run with an estimated maximum likelihood closest to the observed likelihood. I estimated the ancestral effective population size of Hatchet Creek walleye based on the equation of $N_e = \theta\pi\ /2\mu$ (for haploid populations; [50]), with genome-wide nucleotide diversity (θ_π) calculated from STACKS and an assumed mutation rate of 2×10^{-9} per site/generation. Using the parameter estimates associated with the best maximum-likelihood run, I performed 100 bootstrap replicates to estimate the confidence limits in parameter estimation.

3. Results

3.1 GBS Sequencing and SNP Discovery

A total of 69.38 million high-quality reads were generated from Illumina NextSeq sequencing, with an average of 1.16 million reads for each sequenced sample. During SNP discovery, a total of 16,158 SNPs was identified using STACKS after filtering for maf > 0.05, mac > 10, and minimum locus coverage > 0.1 (Fig. 3). Additional stringent filtering steps using VCFtools and SNPRelate packages resulted in a final dataset of 2,782 SNPs (Fig. 3). With Hatchet Creek samples only, I obtained a total of 2,106 SNPs (total concatenated sequence length = 317,028 bp) from STACKS and used this dataset for demography analysis. In order to develop SNP assays for rapid and accurate identification of walleye lineages and various hybrid classes, I identified a dataset of 940 diagnostic SNPs showing fixed genetic differences between northern and southern walleye populations. BAYESCAN failed to identify any SNPs showing evidence of balancing or divergent selection, regardless of the settings of prior odds. The SNP dataset excluding diagnostic loci (1,842 SNPs) was categorized into a high-F_{ST} SNP subset (461, F_{ST}=0.378-0.822), intermediate-F_{ST} subset (921 SNPs, F_{ST}=0.112-0.378), and low-F_{ST} subset (460 SNPs, F_{ST} = 0.001-0.112, see Fig. 4 for F_{ST} distribution). From here on, I use the terms *low*, *intermediate* and *high* SNPs to represent these three subsets.

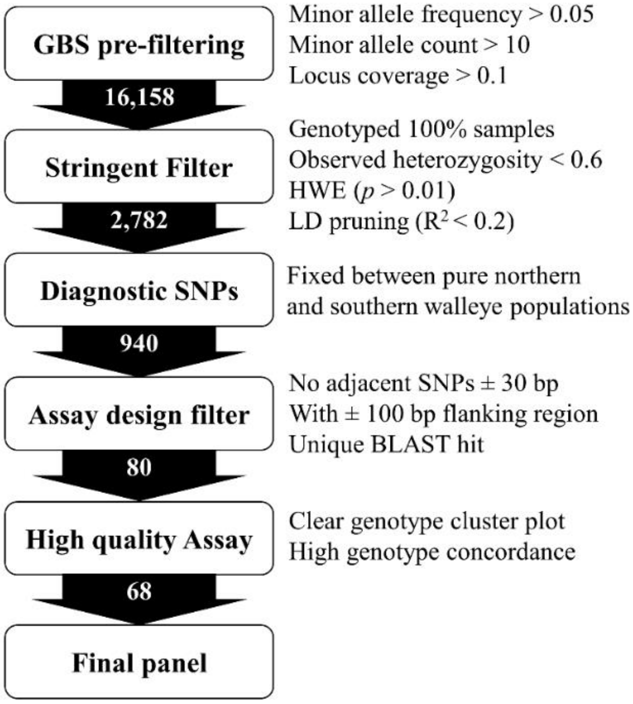
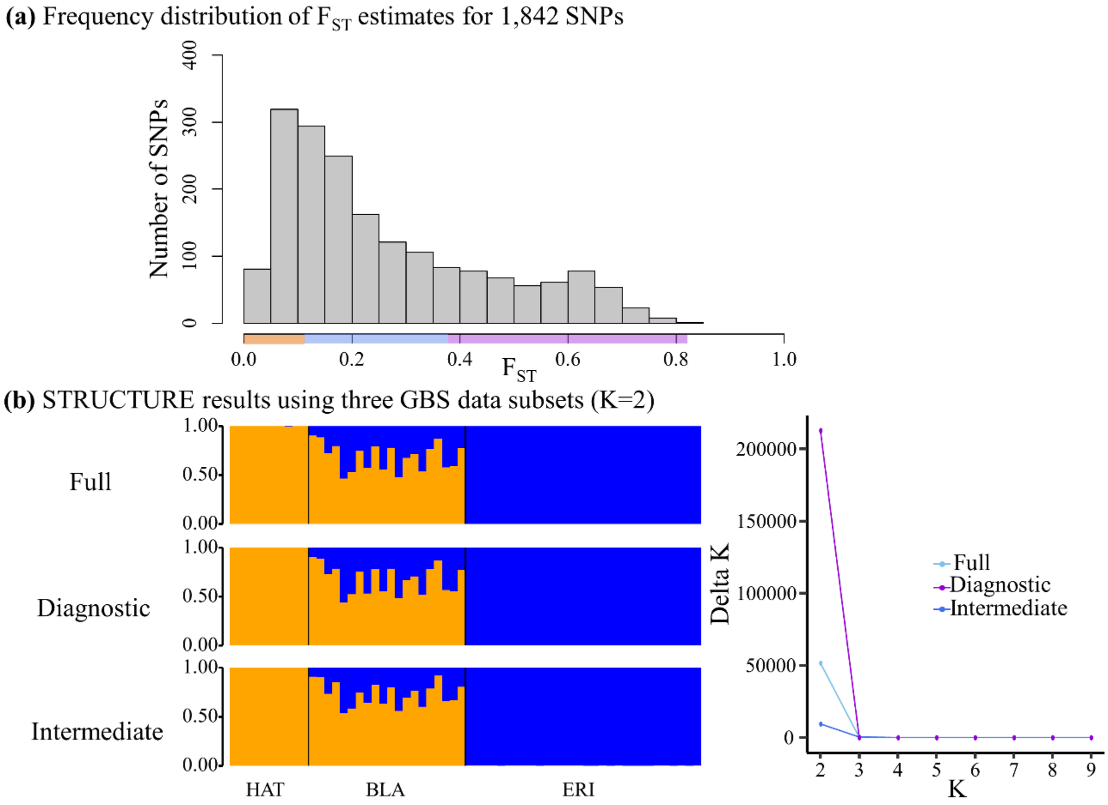


Fig. 3 Workflow demonstrating the steps used in marker identification and selection of markers for diagnostic SNP panel design.


(a) Frequency distribution of F_{ST} estimates for 1,842 SNPs

(b) STRUCTURE results using three GBS data subsets (K=2)

Fig. 4 SNP F_{ST} distribution and population structure results inferred from three GBS data subsets: a) Distribution of locus-specific F_{ST} for GBS SNPs after removal of SNPs fixed between northern and southern walleye. Color bars below the panel indicate the range of F_{ST} for high (461 SNPs, purple), intermediate (921 SNPs, blue) and low (460 SNPs, orange) SNPs; b) STRUCTURE results using K = 2. Three datasets: full, diagnostic, and intermediate were used for STRUCTURE analyses.

3.2 Genetic Diversity and Population Structure Using GBS Data

I measured the genetic diversity of these three populations through the percentage of polymorphic loci (P_o), observed (H_o) heterozygosity and expected heterozygosity (H_e) using the full GBS dataset with 2,782 SNPs. Blackwater Creek had the largest number of variant SNPs (2,652 of 2,782, 95.33%), with Hatchet Creek walleye representing the lowest level of marker polymorphism (444 of 2,782, 15.96%, Table 3). Similarly, the lowest genetic diversity was found in the Hatchet Creek population, with a mean observed heterozygosity of 0.06, suggesting a limited number of founders and/or the presence of drift in this population. I found little evidence of inbreeding in the examined populations, as inbreeding coefficients ranged from -0.05 to 0.002 (Table 3).

Table 3 Number of individuals assayed (*N*), genetic diversity indices and estimated effective population size (N_e) across three walleye populations using GBS data. Diversity indices include the proportion of polymorphic SNPs for each population (*P*), observed heterozygosity (H_o), expected heterozygosity (H_e) and inbreeding coefficient (F_{is}). The full GBS dataset with 2,782 SNPs was used for calculation of diversity indices. The *intermediate* F_{ST} dataset with 921 SNPs was used for estimation of effective population size.

Population	Abbrev.	*N*	*P* (%)	H_o	H_e	F_{is}	N_e (95% CI)
Hatchet Creek	HAT	10	15.96	0.06	0.06	-0.050	10.2 (8.9-11.7)
Blackwater Creek	BLA	20	95.33	0.32	0.32	-0.040	26.1 (25.1-27.1)
Lake Erie	ERI	30	53.34	0.19	0.19	0.002	3103.9 (1068.5-∞)

Pairwise F_{ST} among walleye populations was measured using both the full GBS dataset (2,782 SNPs) and SNP subsets based on locus-specific F_{ST}. Using the full GBS dataset, the obtained F_{ST} estimates ranged from 0.238 (between HAT and BLA) to 0.805 (between HAT and ERI, Table 4). Using SNP subsets, the highest level of F_{ST} was observed when I used *high* SNPs for calculation (ranging from 0.282 between HAT and BLA to 0.841 between HAT and ERI), while pairwise F_{ST} decreased dramatically in the *low* SNP dataset (0.055-0.145). Pairwise F_{ST} estimated from *intermediate* SNPs showed moderate population differentiation, with values ranging from 0.128 (between HAT and BLA) to 0.426 (between HAT and ERI, Table 4). The values generated from *intermediate* SNPs generally mirror the previous F_{ST} estimates from a large-scale walleye genetic divergence study using microsatellites (global F_{ST} of 0.13 ± 0.00; [24]). Therefore, I utilized the *intermediate* subset as neutral SNPs for downstream population structure analyses.

Table 4 Pairwise F_{ST} estimates among walleye populations using GBS SNP data. Three datasets, full (2,782 SNPs), diagnostic (940) and *intermediate* (or neutral, 921) SNPs, were used for F_{ST} calculation. SD is the standard deviation.

	Full		Diagnostic		Intermediate	
	F_{ST}	SD	F_{ST}	SD	F_{ST}	SD
HAT vs. BLA	0.238	0.004	0.323	0.003	0.128	0.006
HAT vs. ERI	0.805	0.005	1.000	0.000	0.426	0.005
BLA vs ERI	0.490	0.005	0.682	0.003	0.236	0.005

A Bayesian model implemented in STRUCTURE was used to assess the best-supported number of high-level genetic clusters (*K*) for walleye populations in the GBS dataset. I examined the STRUCTURE outputs generated from different SNP datasets (full, diagnostic, and *intermediate* F_{ST}); based on the Evanno et al. method [54]; all had the strongest support when $K = 2$ (Fig. 4). In all cases, walleye individuals from Hatchet Creek and Lake Erie represented pure southern and northern walleye alleles, respectively, while walleye individuals sampled from Black Warrior River showed consistent hybridization patterns, suggesting the establishment of a hybrid zone along this river.

3.3 Development and Validation of SNP Panels

Following the previously established protocol for SNP assay design [9, 10, 36], I developed two panels of 40 SNP multiplexes for extensive walleye population genotyping. Detailed information on SNP panels, including the SNP ID, alleles, and primer sequences are listed in Table 5. Among these SNPs, 12 SNPs were excluded from the final assay because of sampling bias or duplicate sequence issues [33, 36]. I observed high concordance of genotype calling between GBS and MassARRAY data, with 99.75% matching genotypes; a similarly high genotype concordance was previously reported for SNP marker development in Florida bass using the same MassARRAY system [36]. I also examined the consistency of MassARRAY genotype calling among technical replicates using a total of 114 individuals, and found 99.88% of genotypes matched across multiple plates.

Table 5 Primer sequence information for the two MassARRAY multiplex panels in walleye.

SNP_ID	SNP Alleles		Primer Sequences
SCTG7180000603687_1802	C/T	PCR1:	ACGTTGGATGTCTCAGAGGAGGTGCGGTTG
		PCR2:	ACGTTGGATGACCCTGTCACCACCGTGAAG
		EXT:	GTACCCCCCACCGTC
SCTG7180000569765_1433	A/G	PCR1:	ACGTTGGATGAATCCCCCTTCTTTGCTCTC
		PCR2:	ACGTTGGATGTTTCTCCGCGAGAAGTTCAG
		EXT:	GCCGGGATCCCTCTT
SCTG7180000611891_1113	G/A	PCR1:	ACGTTGGATGGTTGCAAAACCCAGATACGG
		PCR2:	ACGTTGGATGGATGTGTGTTGTGCAGTGTG
		EXT:	TCGAGCAGAGTGTGA
SCTG7180000682383_540	T/C	PCR1:	ACGTTGGATGTACAATACCTGGCTGTGTGC
		PCR2:	ACGTTGGATGCAGATGAAATGGCAGAAAGG
		EXT:	CCCGTGTACCCTATGA
SCTG7180000535360_4832	T/C	PCR1:	ACGTTGGATGTGTCATTGGGACTCAGGAGG
		PCR2:	ACGTTGGATGTGTGATTTTCCGCAGGACAC
		EXT:	gCTGCCCACTCCCTCTC
SCTG7180000677108_749	A/T	PCR1:	ACGTTGGATGATCTTCTGCCACTCTGACTG
		PCR2:	ACGTTGGATGATAAACTCCACCTGACCCAC
		EXT:	ACCTGACCCACTTGTTA
SCTG7180000536012_760	A/G	PCR1:	ACGTTGGATGTCTCAGAGGATCCAGGAAGG
		PCR2:	ACGTTGGATGGAGATCAGATCTTCTGTGGG
		EXT:	gtGATCAGATGCCGGCT
SCTG7180000846368_789	A/G	PCR1:	ACGTTGGATGAGAGGCTGGATCCAAACTTC
		PCR2:	ACGTTGGATGCAGTAGAGCTGCTTTAAGGG
		EXT:	ccCCTGCCTCTGCCATTT
SCTG7180000553058_6371	G/A	PCR1:	ACGTTGGATGCCTGCTGATCAGTATAACCG
		PCR2:	ACGTTGGATGTTCTGCACATGATCACCGTC
		EXT:	ggaCACCGTCCATGGTGC
SCTG7180000684565_981	C/T	PCR1:	ACGTTGGATGAAAACTACTAGAGGGCCGTG
		PCR2:	ACGTTGGATGGGTGCTCAGTCCATTTAATC
		EXT:	CCTGTGCTTTTCCTCTAAC
SCTG7180000583880_2574*	C/T	PCR1:	ACGTTGGATGGCCTGATATTACACCTGCAC
		PCR2:	ACGTTGGATGCCTTCAGATGCTTGGACTTC
		EXT:	ccttACTTCCACGTCGTGC

SNP_ID	SNP Alleles		Primer Sequences
SCTG7180000844653_153	T/A	PCR1:	ACGTTGGATGAGAGTCACTGCAGCTTTCCA
		PCR2:	ACGTTGGATGTTTCACAACCATCCATCAGC
		EXT:	CCTCAGAGGTTCATTTGAT
SCTG7180000561524_1362	T/A	PCR1:	ACGTTGGATGTCGAGGATCCAAGAAAGGAG
		PCR2:	ACGTTGGATGTGGGTTTCTCCTCCTTTGTG
		EXT:	AAGGAGCATTATGTCTGAA
SCTG7180000533591_7905	G/C	PCR1:	ACGTTGGATGCCAAAGAGCTTTGGCTCTAC
		PCR2:	ACGTTGGATGAGTGTCCAAAATGGATCCTC
		EXT:	tcTGGATCCTCAAACATTCA
SCTG7180000717354_134	A/G	PCR1:	ACGTTGGATGCAGTCTGTGTATCCAGATGC
		PCR2:	ACGTTGGATGTGCTCTCCAGAGTCTCAAAC
		EXT:	ctagGCTGATCAGGATCCTC
SCTG7180000733470_640	G/C	PCR1:	ACGTTGGATGGCCAGCATTTAGCATAGCAC
		PCR2:	ACGTTGGATGTGTACAGTGCAGAGTGCTGG
		EXT:	ggggtTGCTGGGCTAGCTGT
SCTG7180000553198_4339	C/T	PCR1:	ACGTTGGATGCGTCCTCAGGCTCAGTTAAT
		PCR2:	ACGTTGGATGTGATCTACGTTTTACGCCCC
		EXT:	gcgCCATTAAACAAGGAACGC
SCTG7180000611165_3561	C/T	PCR1:	ACGTTGGATGATGTGGCTTTCAGAGGTAGG
		PCR2:	ACGTTGGATGCACACTGAATGACACCTAGC
		EXT:	taAGCGGTACTGCGGTCTAAC
SCTG7180000647964_1188	A/G	PCR1:	ACGTTGGATGACGTTGTAGAAACGTGGTTG
		PCR2:	ACGTTGGATGATAGACCTGCAGTTGCATGG
		EXT:	gggAACGTGGTTGTTTGTAAA
SCTG7180000651972_961*	C/T	PCR1:	ACGTTGGATGCCCATCAAAGGGATCCTTCT
		PCR2:	ACGTTGGATGCAGTGGAAATCACAGAACAG
		EXT:	cAAGGGATCCTTCTTTTCTACT
SCTG7180000635502_1494	T/C	PCR1:	ACGTTGGATGGGATCCTTGCATCTTTGTTG
		PCR2:	ACGTTGGATGAAATGCCCCTTTTTCCCACC
		EXT:	gTGTTGATTTGAGTACCATTTG
SCTG7180000566387_5793	A/G	PCR1:	ACGTTGGATGGACTCAATCCTATATGTGCTG
		PCR2:	ACGTTGGATGCTTCCTACCCTCCAAGTTTG
		EXT:	ccATCCATACCTGTCATAAAAAT
SCTG7180000678642_1484	T/C	PCR1:	ACGTTGGATGTCTCCGTGGGACATTTTAGC
		PCR2:	ACGTTGGATGCCACTCCTCAGTTAACAACC
		EXT:	ccgtGCTCTGTCTTCGCCCTGAA
SCTG7180000562022_5175	G/C	PCR1:	ACGTTGGATGGATCCTTCCTGACATGTCAC
		PCR2:	ACGTTGGATGCAGCTAGCTAGGATTGTTGG
		EXT:	CTAGGATTGTTGGTTGTTCATTT
SCTG7180001057026_4442	G/A	PCR1:	ACGTTGGATGATTTGGGTGTACGAGGAGTC
		PCR2:	ACGTTGGATGCCTGGGAGACCTTGTAGTAA
		EXT:	cttcTTCATATTGGCTGGATATCT
SCTG7180000542333_8207	G/C	PCR1:	ACGTTGGATGCAAGGGAGACCATGAGTTAC
		PCR2:	ACGTTGGATGAGGACACTGCAGTATTAGAG

SNP_ID	SNP Alleles		Primer Sequences
		EXT:	aAGTATTAGAGTGTTAACACTTAG
SCTG7180000557521_4509	T/C	PCR1:	ACGTTGGATGCTCTACAGTCTCCCCCTAAG
		PCR2:	ACGTTGGATGTGTCGTATGTACCTGTCCAC
		EXT:	ccccTACCTGTCCACCTGCAAAGAT
SCTG7180000545873_3318	C/T	PCR1:	ACGTTGGATGCCACCTCCCACAGATCTAAC
		PCR2:	ACGTTGGATGCCTGAACTCCCCAATTCTTG
		EXT:	CTCCCACAGATCTAACTGTAAATTG
SCTG7180000541887_17531	C/T	PCR1:	ACGTTGGATGTCCTTCTGCCCCTCAGAAAG
		PCR2:	ACGTTGGATGGAAAGCTAAAACCGCAGCTC
		EXT:	gggaaTCAGAAAGGATCCTTTGTAC
SCTG7180000565869_3163	T/C	PCR1:	ACGTTGGATGTCTCCAGATACCAGTGGAAC
		PCR2:	ACGTTGGATGTTCTGTGCTAACTGGAGGTC
		EXT:	ggcaAGACCATCGGATCAGGTAGCG
SCTG7180000543610_8300*	C/T	PCR1:	ACGTTGGATGCTCAGCTATTCTCTCTTTCAC
		PCR2:	ACGTTGGATGCCCAGGATCCACCAAAGTTC
		EXT:	gcatCTATTCTCTCTTTCACGGTAAA
SCTG7180000611699_752	A/G	PCR1:	ACGTTGGATGTTCATGGATCCAGTGGAAGC
		PCR2:	ACGTTGGATGAACAAGTTTCAAGTCTGAGC
		EXT:	cccccATGTCAGCATGATATAAACTT
SCTG7180000549979_9596	A/C	PCR1:	ACGTTGGATGTGCAGCTGTGGCCTTTTAAG
		PCR2:	ACGTTGGATGAGAAAGGTCTGGAAAGCCAC
		EXT:	tatgAGTACCACAACTATGTAAATAC
SCTG7180000556688_2595	A/G	PCR1:	ACGTTGGATGGTTCTTAGAGAGGATCCCTG
		PCR2:	ACGTTGGATGATGCAGTGGCAGGTAATGTG
		EXT:	acagaAGGTAATGTGGAGCTTTTTGA
SCTG7180000612634_2040	T/C	PCR1:	ACGTTGGATGGGGAGGAAAGAAGGAAAAG
		PCR2:	ACGTTGGATGGATCCAGCACAGTGCAAGAG
		EXT:	gagagCAGTGCAAGAGTTAAATTGAC
SCTG7180000541289_5468*	C/G	PCR1:	ACGTTGGATGGCCATTGGCATTTATTGGAC
		PCR2:	ACGTTGGATGTTCTAAAAAACTGGATCCC
		EXT:	ggggAACTGGATCCCATAATGCAACTT
SCTG7180001057293_1417	A/G	PCR1:	ACGTTGGATGGCATGTTGTGCCTTTTCTTG
		PCR2:	ACGTTGGATGAGCTTCCACAAAGTGTTCGC
		EXT:	tttcTCCACAAAGTGTTCGCAAGTCTTC
SCTG7180000612720_3290	A/T	PCR1:	ACGTTGGATGTTTTTAGCGGGTCTCAGGTC
		PCR2:	ACGTTGGATGAGATCAGAGCTGGAGTGAAC
		EXT:	accTATCTAATACCCTGGTGGATCCAAA
SCTG7180000725834_2641	C/T	PCR1:	ACGTTGGATGAAACAACAACTGGGCACAGG
		PCR2:	ACGTTGGATGGGTTAATAGACATAACTGG
		EXT:	GGTTAAACTGATTTTGAAAATCTATTAG
SCTG7180000628127_1356	A/T	PCR1:	ACGTTGGATGGAGTGGGATCCTCCTTTTTC
		PCR2:	ACGTTGGATGGGGTTGTAGTTGGGAACCAC
		EXT:	tatagCAAAGGAGCAAGATAATTTATTA
SCTG7180000629308_1771	T/A	PCR1:	ACGTTGGATGTTCAGTCACATGCACACACC

SNP_ID	SNP Alleles		Primer Sequences
		PCR2:	ACGTTGGATGGAAATAGGATCCAGAGTCCC
		EXT:	ACACACCACACCAAT
SCTG7180000571636_5033	T/C	PCR1:	ACGTTGGATGCAGAGTGTGAAAGGATTGGC
		PCR2:	ACGTTGGATGCCTGAATGAAGAAGCTGTCG
		EXT:	GCTGTCGTTTGGAGC
SCTG7180000576929_1141	G/A	PCR1:	ACGTTGGATGAGAAATCCTGAGAGCTGCTG
		PCR2:	ACGTTGGATGTTCAGTGAGCCTCTCCAAAG
		EXT:	cGTCCCGCCTGTCTCT
SCTG7180000560505_2795	A/G	PCR1:	ACGTTGGATGCAACAGACAGGAAGAACTGG
		PCR2:	ACGTTGGATGGTAGATTTGAGCTTCTTGGC
		EXT:	GCTTCTTGGCCATCTT
SCTG7180000551118_2381	T/G	PCR1:	ACGTTGGATGGGAGTCATTATTGAAAGCGG
		PCR2:	ACGTTGGATGCGCCACATACTGCCTATAAG
		EXT:	AAGCGGAGATACGTTC
SCTG7180000554096_5503	A/C	PCR1:	ACGTTGGATGCACCTTTCTGTACAGTGATG
		PCR2:	ACGTTGGATGTTTTTCAGGATCGTTGCCCG
		EXT:	agCGTTGCCCGACGCAT
SCTG7180000534531_11968	C/G	PCR1:	ACGTTGGATGGCGTGTGTACAGTATGTGTG
		PCR2:	ACGTTGGATGTCCTTTTTCCTCCCTGGAAG
		EXT:	ggGGAAGCCCAAATGGA
SCTG7180000544033_8226	G/C	PCR1:	ACGTTGGATGCAACGAGACATTTCCCAACC
		PCR2:	ACGTTGGATGTAGGGCAGAACTGAACTAGC
		EXT:	ctATCCACTTTTCAGCCT
SCTG7180000546758_2639	A/G	PCR1:	ACGTTGGATGAGATGCAAACACTGGCTTGG
		PCR2:	ACGTTGGATGACAGAAAGTACAGCAGTGGC
		EXT:	GCCGATACCCAAATATCT
SCTG7180000556590_1391	A/G	PCR1:	ACGTTGGATGCAGCAAAAGCCATACCTGTG
		PCR2:	ACGTTGGATGTCTTGGATCCATCTGTCCTG
		EXT:	cTTTTGCCTCATTGCTTTA
SCTG7180000540968_9812	T/C	PCR1:	ACGTTGGATGTGCCTGTCTATTCACATGGG
		PCR2:	ACGTTGGATGTTAGATGGTAAGCGGCGCTC
		EXT:	gactCTCTGCCTGCGTACC
SCTG7180000536953_4769	T/G	PCR1:	ACGTTGGATGGATCCACGGCAGTGATGTTG
		PCR2:	ACGTTGGATGCTGACTCCCCTCATTTTCTG
		EXT:	TACAGACATTTTGACAGGA
SCTG7180000532744_5405	T/C	PCR1:	ACGTTGGATGGGCCAAAGCCAAAGAAAAGG
		PCR2:	ACGTTGGATGTCCGGGATGGGATAGGAAAC
		EXT:	ctcccCCTGGATCCCCCTCG
SCTG7180000736955_246	A/G	PCR1:	ACGTTGGATGAGACACCTAGATTGTGAGGC
		PCR2:	ACGTTGGATGGGTTCACTGACTCACTCTTG
		EXT:	GTCTCAATCCTATGTCAGTA
SCTG7180000531956_2425*	T/G	PCR1:	ACGTTGGATGTTCACATCTACACGGTGTCC
		PCR2:	ACGTTGGATGTCACCTTCTCAGGATCCATC
		EXT:	ggccGGATCCATCCGAAACA

SNP_ID	SNP Alleles		Primer Sequences
SCTG7180000578506_4179	G/A	PCR1:	ACGTTGGATGCAGTATCAAGGCTATGGGTG
		PCR2:	ACGTTGGATGAAATAGGATCCTTGCCTGCC
		EXT:	cccccCCTGCCTGATACACAA
SCTG7180000728197_2357	A/G	PCR1:	ACGTTGGATGTCCTGATTGCCTCTTCTCAC
		PCR2:	ACGTTGGATGAGATAGCAAGACGCTGACAC
		EXT:	ccTCACAATCGGATTCCACTC
SCTG7180000630055_1140*	T/C	PCR1:	ACGTTGGATGCCGTTAGTTGGTTAGTGTTG
		PCR2:	ACGTTGGATGAGCTAAGCCGTGTAAGTGAC
		EXT:	atCCGTGTAAGTGACTGCTGC
SCTG7180000752917_410	A/G	PCR1:	ACGTTGGATGGGACTGGATCCGAATATTTG
		PCR2:	ACGTTGGATGCGTTTGTGCAAATTTCAGCC
		EXT:	ccccaAAATTTCAGCCTGCATG
SCTG7180000588781_1928*	A/G	PCR1:	ACGTTGGATGCAGCTGCTGACAGACTGTTG
		PCR2:	ACGTTGGATGATTACAGGGTTGGATCCAAG
		EXT:	TGCTACACATTCAGTACATTTC
SCTG7180000572504_3166	C/G	PCR1:	ACGTTGGATGCCTATGCACTGTTAGGAGAC
		PCR2:	ACGTTGGATGTCATAAAGGATCCCAGCGTG
		EXT:	TCATCTAAGACAGGATGAATCA
SCTG7180000578534_6039	C/T	PCR1:	ACGTTGGATGTTTAGACCTCTCGCTTGGAC
		PCR2:	ACGTTGGATGTGTGTCCTGCTATGTTCTCC
		EXT:	tctcTGTTCTCCCCTTTTCATAT
SCTG7180000579975_3985	T/C	PCR1:	ACGTTGGATGAACTGTGTCTCTCTCACCAG
		PCR2:	ACGTTGGATGGCGTTGTCTTTAACACAAGG
		EXT:	agccgCAGGTCAGTCTCGGTCGA
SCTG7180000531982_14467	C/A	PCR1:	ACGTTGGATGCATGCAGAGTGGAAGACTTG
		PCR2:	ACGTTGGATGCCGTCTTGAGCACAAGCAAC
		EXT:	gtgtAAGACTTGTCTTATCAGGA
SCTG7180000566957_2312	T/C	PCR1:	ACGTTGGATGGCAGTAGTTGCATTTGTCAC
		PCR2:	ACGTTGGATGTCCGGGAATCTCAGTGAAAG
		EXT:	cttcGTTGCATTTGTCACCATCTT
SCTG7180000558940_1956*	G/A	PCR1:	ACGTTGGATGAGGTCATAGCAGGAACAGAG
		PCR2:	ACGTTGGATGATAAATACAGACTTCTGTG
		EXT:	cATCTAAACCTCCAAGTATACATG
SCTG7180000539586_6310	A/T	PCR1:	ACGTTGGATGAGTTCAGCAGCTAACATAGG
		PCR2:	ACGTTGGATGCTGGATCCTCTCTTACCTTG
		EXT:	ggtgGATTAATTGCACCAGGATTT
SCTG7180000574637_2241	T/C	PCR1:	ACGTTGGATGTGAGTTGTGCCTAAAGTGCC
		PCR2:	ACGTTGGATGTTGCCAGCAAAACCACGGAG
		EXT:	ttgacACCACGGAGGATCCCTTGCT
SCTG7180000679954_1362*	T/G	PCR1:	ACGTTGGATGGGGTTAGGGTTAGGTGTTTT
		PCR2:	ACGTTGGATGACTGTGTCACAGCATGGTTG
		EXT:	AACGCAGTACTTAAATTTTTTTAGT
SCTG7180000536390_6120	T/C	PCR1:	ACGTTGGATGGGACAGGACATTTGTCATAC
		PCR2:	ACGTTGGATGCCCTATCATTCTGCCATGAG

SNP_ID	SNP Alleles		Primer Sequences
		EXT:	ggtggGAAGTCAGTGCGCACATAAA
SCTG7180000566569_462	A/T	PCR1:	ACGTTGGATGCATAACTCACTGGATCCCTG
		PCR2:	ACGTTGGATGGGGTTCTTCACCTTGTTACG
		EXT:	tgtcTCCCTGACCTTTCAGTAACACA
SCTG7180000537339_2135	G/A	PCR1:	ACGTTGGATGTGTGTGTAATGCTAGAGGAG
		PCR2:	ACGTTGGATGGCTTTTGGACAGGAACTATG
		EXT:	aaaagGTGTCATATAAGCTTTCACAG
SCTG7180000543744_7319	A/G	PCR1:	ACGTTGGATGTACGCCAGCAAGTTGAAAGG
		PCR2:	ACGTTGGATGCCAGGTCAAGGGAAACAATC
		EXT:	ggaAAAGGTAACCTTAATCTGAGCCT
SCTG7180000728936_897	T/C	PCR1:	ACGTTGGATGCACCTCAGTTCTCAACCCAG
		PCR2:	ACGTTGGATGACTCACTCATCCTCAGCTTG
		EXT:	acCCTGCTTATTGATGACAAGACCCTT
SCTG7180000557112_7099	T/C	PCR1:	ACGTTGGATGCCATTTGTAGCTGTAGCGTG
		PCR2:	ACGTTGGATGACTTAGCGCAGCTTTCATAG
		EXT:	cgtatCTGTAGCGTGCAACCACAACTT
SCTG7180000648204_2443*	C/A	PCR1:	ACGTTGGATGGAGAAGGATCCCTCCTCAAG
		PCR2:	ACGTTGGATGAGGAAAACCTCCCAAAAAAC
		EXT:	ggacAATGAAGAAATCTTCAGGAAAAC
SCTG7180000629677_2087	A/G	PCR1:	ACGTTGGATGATCCTATTGTCTCATTCGCC
		PCR2:	ACGTTGGATGTTTGCCAATGACAGCCCTTC
		EXT:	tttttCCAATGACAGCCCTTCAACCTTC
SCTG7180000532610_14667	C/T	PCR1:	ACGTTGGATGAAACGAGGGAGAATGTGTGC
		PCR2:	ACGTTGGATGGATCCCCTGTCCATTAGTCC
		EXT:	CCCCTGTCCATTAGTCCAGCTTTTATTC
SCTG7180000649874_427*	A/C	PCR1:	ACGTTGGATGTACTTTGGATCCACTACCAC
		PCR2:	ACGTTGGATGTCCATCCAGTGTAGTCATGC
		EXT:	ataccTGCTACAGGGACTCAACCAACAC
SCTG7180000581699_1488*	A/T	PCR1:	ACGTTGGATGCATCAGGTCAATGACATGG
		PCR2:	ACGTTGGATGCCTAATAGGAACAGAAAAGAG
		EXT:	cgggTTTTGTTACACTGTACAATTACAA

* Excluded from SNP panels for genotyping

3.4 Genetic Structure of 23 Walleye Populations Using the SNP Assay

Using the 68-SNP assay, a total of 545 additional walleye individuals across 23 populations were genotyped for diagnostic SNP validation, extensive population structure analysis, and hybrid classification. I successfully genotyped these samples with a high genotyping rate (average 98.98%). Using STRUCTURE and the Evanno et al. (2005) [54] criterion, I found the optimal number of genetic clusters was *K*=2, which was consistent with the results from GBS datasets (Fig. 5b). After removing the southern walleye lineages in the STRUCTURE analysis, I observed two major genetic clusters: one formed by walleye individuals from the Great Lakes (HUR, ERI, MUS, THU), upper Mississippi River (MIL) and Little Tennessee River (NAN and FON): and another one represented by individuals from eastern highlands (e.g., FOS and ROC, Fig. 5b). When the non-parameter-based DAPC method was used for population clustering, four genetic clusters received the strongest support (Fig. 5a), potentially due to the existence of eastern highlands walleye populations and hybrid populations from Black Warrior River. After removing the hybrid individuals from the Black Warrior River, three major genetic clusters were observed among examined walleye individuals (Fig. 5a).

Population differentiation analyses using Hudson's F_{ST} revealed the same divergent patterns, as the largest level of genetic differentiation was found between the southern (HAT, WHI, and TOM) and upper northern groups (including NAN and FON populations from the Little Tennessee River, Fig. 6). Walleye populations in eastern highlands drainages (e.g., FOS and ROC) were genetically differentiated from all other walleye groups.

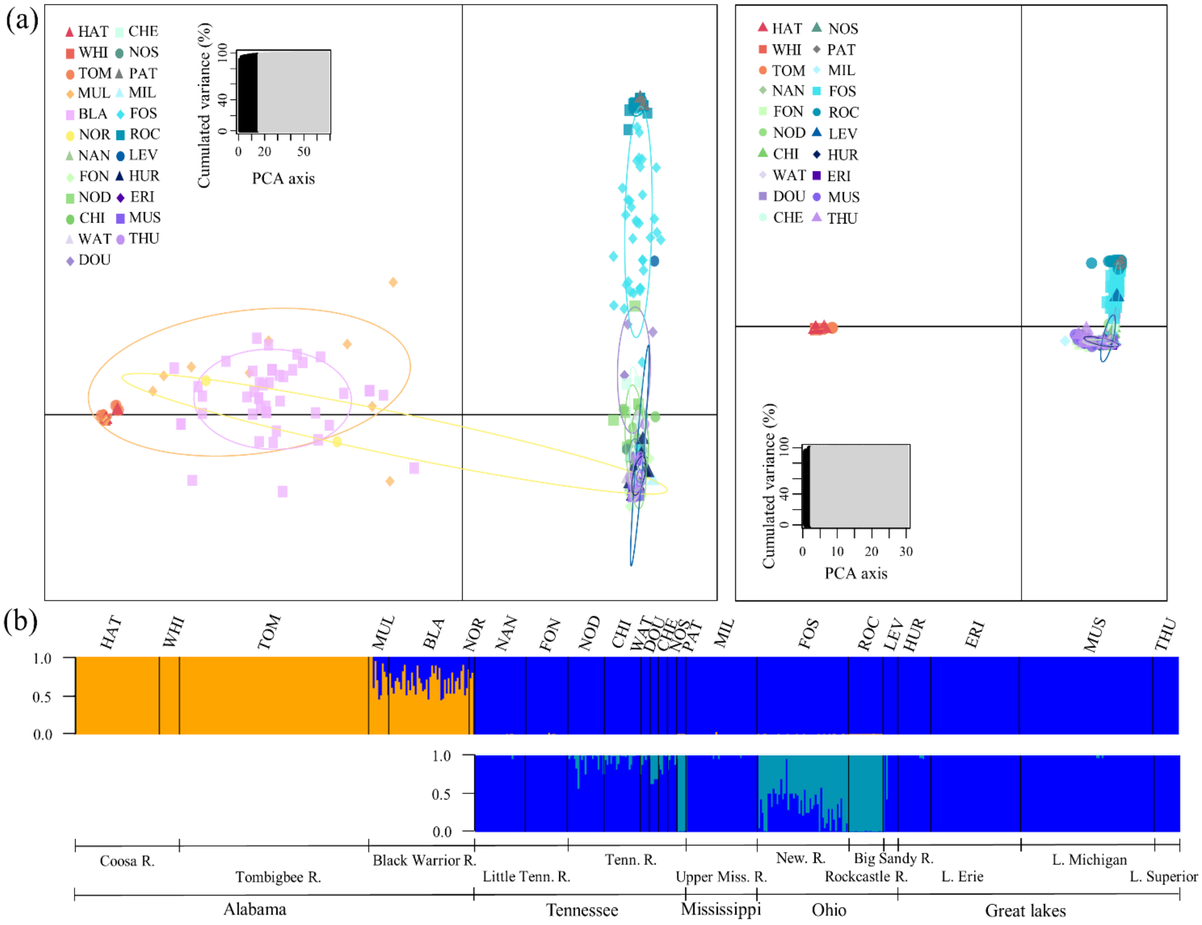
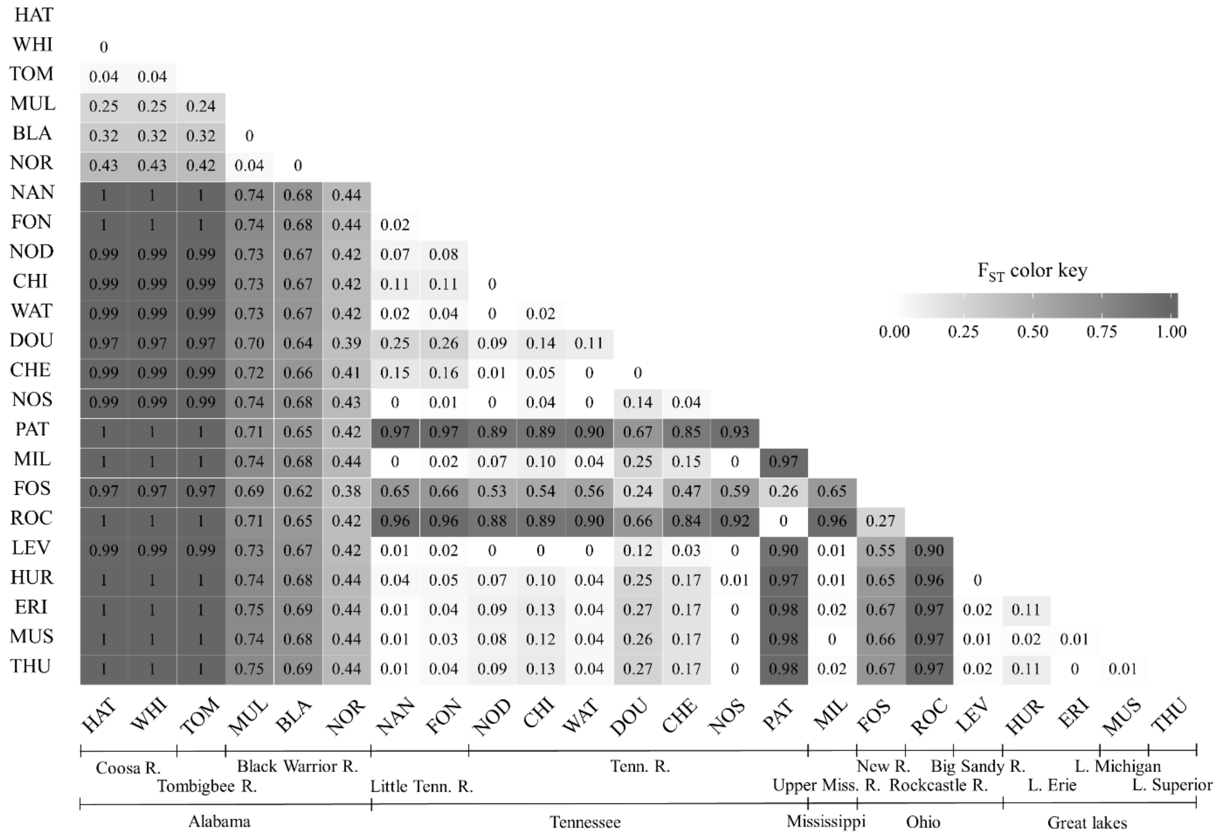


Fig. 5 Population structure results inferred from DAPC and STRUCTURE using 68-SNP genotyping data. a) Scatterplot output from DAPC for the genetic cluster of walleye individuals with (left) and without (right) hybrid populations; b) STRUCTURE result using $K = 2$ for all (upper) and northern (bottom) walleye individuals.



	HAT	WHI	TOM	MUL	BLA	NOR	NAN	FON	NOD	CHI	WAT	DOU	CHE	NOS	PAT	MIL	FOS	ROC	LEV	HUR	ERI	MUS	THU
HAT																							
WHI	0																						
TOM	0.04	0.04																					
MUL	0.25	0.25	0.24																				
BLA	0.32	0.32	0.32	0																			
NOR	0.43	0.43	0.42	0.04	0																		
NAN	1	1	1	0.74	0.68	0.44																	
FON	1	1	1	0.74	0.68	0.44	0.02																
NOD	0.99	0.99	0.99	0.73	0.67	0.42	0.07	0.08															
CHI	0.99	0.99	0.99	0.73	0.67	0.42	0.11	0.11	0														
WAT	0.99	0.99	0.99	0.73	0.67	0.42	0.02	0.04	0	0.02													
DOU	0.97	0.97	0.97	0.70	0.64	0.39	0.25	0.26	0.09	0.14	0.11												
CHE	0.99	0.99	0.99	0.72	0.66	0.41	0.15	0.16	0.01	0.05	0	0											
NOS	0.99	0.99	0.99	0.74	0.68	0.43	0	0.01	0	0.04	0	0.14	0.04										
PAT	1	1	1	0.71	0.65	0.42	0.97	0.97	0.89	0.89	0.90	0.67	0.85	0.93									
MIL	1	1	1	0.74	0.68	0.44	0	0.02	0.07	0.10	0.04	0.25	0.15	0	0.97								
FOS	0.97	0.97	0.97	0.69	0.62	0.38	0.65	0.66	0.53	0.54	0.56	0.24	0.47	0.59	0.26	0.65							
ROC	1	1	1	0.71	0.65	0.42	0.96	0.96	0.88	0.89	0.90	0.66	0.84	0.92	0	0.96	0.27						
LEV	0.99	0.99	0.99	0.73	0.67	0.42	0.01	0.02	0	0	0	0.12	0.03	0	0.90	0.01	0.55	0.90					
HUR	1	1	1	0.74	0.68	0.44	0.04	0.05	0.07	0.10	0.04	0.25	0.17	0.01	0.97	0.01	0.65	0.96	0				
ERI	1	1	1	0.75	0.69	0.44	0.01	0.04	0.09	0.13	0.04	0.27	0.17	0	0.98	0.02	0.67	0.97	0.02	0.11			
MUS	1	1	1	0.74	0.68	0.44	0.01	0.03	0.08	0.12	0.04	0.26	0.17	0	0.98	0	0.66	0.97	0.01	0.02	0.01		
THU	1	1	1	0.75	0.69	0.44	0.01	0.04	0.09	0.13	0.04	0.27	0.17	0	0.98	0.02	0.67	0.97	0.02	0.11	0	0.01	

Fig. 6 Pairwise Hudson's F_{ST} estimate for walleye populations using the 68-SNP assay.

3.5 Hybrid Classification

To evaluate walleye hybridization status in the Black Warrior River, I applied a novel framework that combined both a manual and NEWHYBRIDS methods for hybrid classification. The hybrid analyses were conducted using 300 fixed SNPs from the GBS dataset and 68 SNPs from MassARRAY assays. Before the analyses, I tested the performance of SNPs in assigning simulated individuals to 12 hybrid classes using NEWHYBRIDS. Using three random subsets of 300 SNPs, the average accuracy was 100% for F_1, F_2, and first-generation backcrosses (B_x-), 99.3% for second-generation backcrosses (B_x2-), 94.0% for third-generation backcrosses (B_x3-), and 99.4% for fourth-generation (B_x4-) backcrosses (Fig. 7). For simulation using the 68-SNP assay data, a correct assignment was made for 100% of F_1 and F_2 hybrids, 95.0% of first-generation backcrosses (B_x-), 89.0% of second-generation backcrosses (B_x2-), 62.0% of third-generation backcrosses (B_x3-), and 77.0% of fourth-generation backcrosses (B_x4-). The mis-assigned individuals were composed of hybrids from later backcross generations, potentially due to the close genotype probabilities among later generation categories (e.g., B_x3 mis-assigned as B_x2 or B_x4).

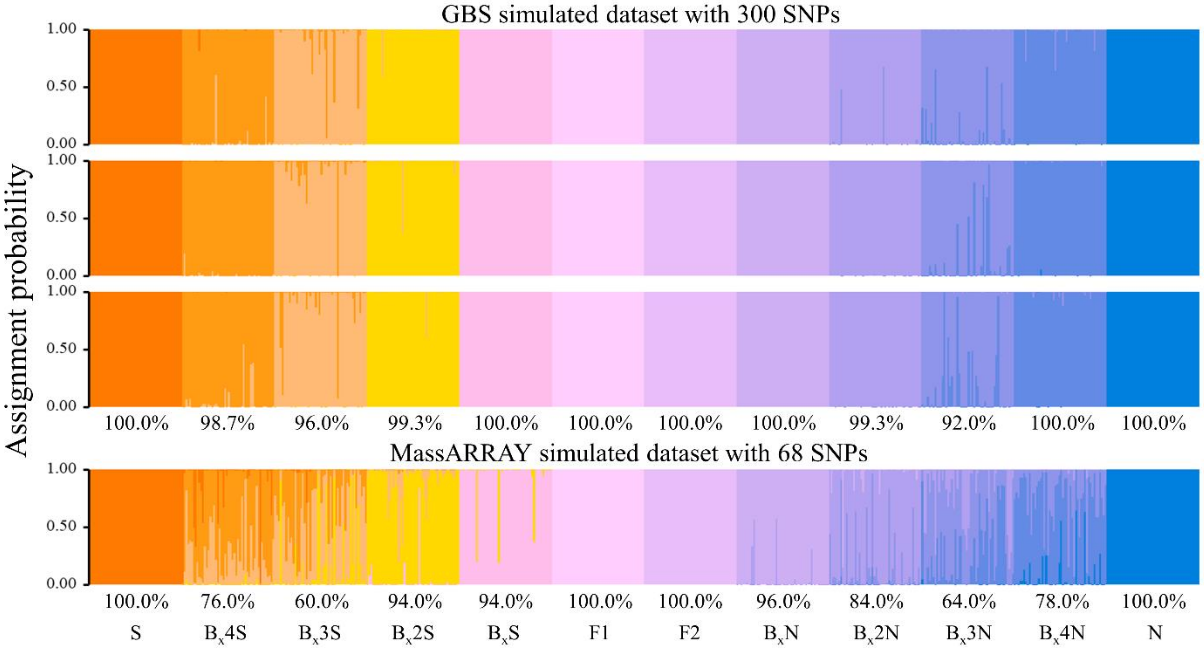


Fig. 7 NEWHYBRIDS simulation analyses based on three GBS datasets with 300 randomly selected diagnostic SNPs and MassARRAY dataset with 68 SNPs. Twelve hybrid categories (listed at the bottom of the figure; defined in Table 2) were set as “Jeffreys-like priors”. Values at the bottom of the figure show the accuracy of assignment to each simulated hybrid category.

I also measured the assignment concordance between GBS and MassARRAY results in 20 hybrid walleye individuals. From the comparison, all but two hybrid assignments gave congruent classifications (Table 6). The two mis-assignments were due to rare northern walleye alleles not captured by MassARRAY SNPs. Using the same classification methods, hybridization between northern and southern walleye was detected in additional walleye sampled from the Black Warrior River. As shown in Table 7, advanced stages of hybridization were the most commonly observed categories among the 55 hybrid individuals, with early-generation hybrids (B_xS and B_xN) making up only 7.3% of total hybrids sampled (Table 7). Later-generation hybrids (B_x2, B_x3, F_x, F_2) were dominated by the F_2 (54.9%) and F_xS categories (29.4%). I observed some genetically pure walleye in Black Warrior River sections, with 2 out of 11 fish identified as pure southern walleye in the Mulberry Fork, and 1 out of 3 fish as pure northern walleye in the North River.

Table 6 Comparison of hybrid assignments between GBS and MassARRAY datasets. A total of 20 walleye individuals from Black Warrior River were used for the analysis.

Sample ID	GBS				MassARRAY			
	Hybrid	S(AA)	H(AB)	N(BB)	Hybrid class	S(AA)	H(AB)	N(BB)
WEE_24	Bx2S	0.797	0.203	0	Bx2S	0.779	0.221	0
WEE_62	Bx2S	0.770	0.230	0	Bx2S	0.779	0.221	0
WEE_63	F2	0.467	0.504	0.029	F2	0.441	0.544	0.015
WEE_64*	FxS	0.610	0.372	0.018	Bx2S	0.721	0.279	0
WEE_65	F2	0.014	0.864	0.121	F2	0.015	0.868	0.118
WEE_66	F2	0.312	0.406	0.282	F2	0.235	0.471	0.294
WEE_67	FxS	0.521	0.432	0.047	FxS	0.485	0.456	0.059
WEE_68	F2	0.256	0.544	0.200	F2	0.254	0.567	0.179
WEE_69	FxS	0.637	0.294	0.069	FxS	0.559	0.338	0.103
WEE_70	F2	0.280	0.536	0.184	F2	0.279	0.515	0.206
WEE_71	FxS	0.614	0.351	0.034	FxS	0.618	0.368	0.015
WEE_72	F2	0.229	0.524	0.247	F2	0.250	0.574	0.176
WEE_73	F2	0.363	0.610	0.027	F2	0.412	0.559	0.029
WEE_74	F2	0.427	0.554	0.019	F2	0.485	0.500	0.015
WEE_75	F2	0.260	0.532	0.208	F2	0.309	0.559	0.132
WEE_76	FxS	0.642	0.303	0.054	FxS	0.691	0.265	0.044
WEE_77	Bx2S	0.752	0.248	0	Bx2S	0.794	0.206	0
WEE_78	F2	0.209	0.683	0.108	F2	0.147	0.735	0.118
WEE_79	F2	0.298	0.522	0.180	F2	0.353	0.500	0.147
WEE_80*	FxS	0.563	0.418	0.019	Bx2S	0.676	0.324	0

* Individuals assigned differently using GBS and MassARRAY data

Table 7 Summary of the hybrid analyses on Black Warrior River walleye samples based on manual and NEWHYBRIDS assignment. Twelve hybrid categories (defined in Table 2) were set as "Jeffreys-like priors".

	S	B_x4S	B_x3S	B_x2S	B_xS	F1	F2	B_xN	B_x2N	B_x3N	B_x4N	N	F_xN	F_xS	Total
Mulberry Fork															
N	2	0	1	0	0	0	5	0	0	0	0	0	0	3	11
%	18.2	0	9.1	0	0	0	45.5	0	0	0	0	0	0	27.3	
Blackwater Creek (BLA)															
N	0	0	0	7	3	0	23	0	0	0	0	0	0	11	44
%	0	0	0	15.9	6.8	0	52.3	0	0	0	0	0	0	25	
North River (NOR)															
N	0	0	0	0	1	0	0	0	0	0	0	1	0	1	3
%	0	0	0	0	33.3	0	0	0	0	0	0	33.3	0	33.3	

3.6 Historical Demographic Analyses of Hatchet Creek Walleye

The N_e estimates using *intermediate* SNPs showed that walleye in Hatchet Creek and Blackwater Creek had relatively low effective population sizes compared to the walleye population sampled from Lake Erie (Table 3). Given the relatively low genetic diversity and effective population size in Hatchet Creek samples, I assessed four demographic models to reconstruct the population's demographic history (Fig. 2). The continuous decline model (M2) was the best-supported model, regardless of assigned mutation rates (Fig. 2 and Table 8). Based on this model, I obtained a coalescent diploid N_e estimate of 7 for the current Hatchet Creek walleye population, with an average population growth rate of 1.33E^{-05} (Table 9). By combining the values for mutation rate (2×10^{-9} per site/generation) with nucleotide diversity ($\theta_\pi = 0.166\%$, estimated from STACKS), I estimated a long-term haploid effective population size that approximated 415,000. Given that the long-term harmonic mean of N_e corresponds to $2N_e$ generations (with more weight given to recent generations, [55]), the estimated ancestral haploid N_e can be traced back to ~800,000 generations ago.

Table 8 Parameter setting in fastsimcoal and model selection using AIC.

Mutation Rate	Model	MaxEstLhood	N of parameters	AIC	ΔAIC	AIC's
2.0E-09	M1	-11871.43	1	23744.87	9342.92	0
	M2	-7197.98	3	14401.95	0	1
	M3	-7673.23	3	15352.47	950.52	0
	M4	-7673.18	4	15354.35	952.40	0
3.5E-09	M1	-11360.79	1	22723.57	8319.73	0
	M2	-7198.92	3	14403.84	0	1
	M3	-7576.67	3	15159.34	755.51	0
	M4	-7568.22	4	15144.43	740.59	0
2.5E-08	M1	-9583.87	1	19169.74	4766.57	0
	M2	-7198.59	3	14403.18	0	1
	M3	-7556.16	3	15118.32	715.14	0
	M4	-7556.03	4	15120.06	716.89	0

Table 9 Parameter estimates based on the continuous decline model (M2), the preferred model in the FASTSIMCOAL analysis. Diploid coalescent effective population sizes are provided in the table.

Parameter	Average estimate	Minimum	Maximum
Current population size	7	1	28
Population growth rate	1.33E-05	1.10E-05	1.59E-05

4. Discussion

Walleye is an ecologically and economically valuable freshwater species in North America that is threatened by overexploitation and genetic risks from anthropogenic activities. This study is the first to utilize genotype-by-sequencing and a novel SNP assay to characterize population genetic structure and hybridization among northern and southern walleye populations. Using 2,782 SNPs, I confirmed the genetic distinctness of Mobile River Basin walleye and identified an anthropogenic hybrid zone that likely resulted from the restocking of northern walleye into the Black Warrior River system of Alabama. I also found strong evidence of a historical declining population trend with reduced genetic diversity and effective population size in a pure southern walleye population from Hatchet Creek, Alabama. Using a newly developed 68-SNP assay, I extensively genotyped 23 walleye populations and described broad-scale genetic structure, distinguishing three groups comprised of the Mobile River Basin, Great Lakes/upper Mississippi, and eastern highlands populations. I have shown that a suite of 68 SNPs can collectively classify advanced-generation hybrids between northern and southern individuals, which will be useful for conservation and protection of putatively locally adapted stocks in the Mobile River Basin.

4.1 Genetic Divergence Between Southern and Other Walleye Groups

A major goal of this study was to characterize the genetic differentiation between northern and southern walleye groups. In the current study, the unique genetic pattern of southern Mobile River Basin walleye was characterized using several approaches. First, I determined that 940 (33.79%) of the 2,782 GBS SNPs were fixed between Hatchet Creek (HAT) and Lake Erie (ERI) walleye populations, indicating genome-wide divergence and lack of gene flow between the two groups. For comparison, the ratio of diagnostic walleye SNPs identified in this study was remarkably higher than the level of interspecific difference (9.2%, 675 of 7,346 loci) observed between Russian (*Acipenser gueldenstaedtii*) and Persian (*A. persicus*) sturgeon [56]. These diagnostic loci can be useful for investigating selection and adaptive evolution, heterogeneous genome divergence, and intraspecific introgression or hybridization [57, 58].

I also examined population structure and differentiation using both GBS and MassARRAY data, and found three major lineages comprised of walleye individuals from the Mobile River Basin, eastern highlands drainages, and upper northern regions, with the Mobile River Basin lineage strikingly different from other walleye populations (Fig. 4, 5 and 6, Table 4). This pattern was still supported when I excluded hybrid individuals from the Black Warrior River in the DAPC cluster analyses (Fig. 5). The distinctiveness of this walleye lineage may be due to their long undisturbed history of local adaption and independent evolution in this isolated southern river system [19, 59]. A similar genetic structure was previously reported for southerly populations of yellow perch (*Perca flavescens*), a species with a similar native range and life history characteristics to walleye [60]. Lastly, the GBS data showed that the walleye population in Hatchet Creek had lower genetic diversity and estimated effective population size compared to the walleye populations sampled from Lake Erie and Blackwater Creek (Table 3), reflecting the long-term isolation and potential historical population decline in pure southern groups [24].

4.2 SNP Assay Resolution for Population Structure

Although the 68-SNP assay is able to identify pure and hybrid southern walleye in a rapid and accurate manner, the limited number of representative populations in the GBS data and utilization of fixed markers may restrict the assay application for resolving fine-scale genetic structure among extensive walleye populations. For example, I observed three major genetic clusters among examined walleye individuals using the SNP assay (Fig. 5), which only covers the broad-scale structure patterns previously identified from mtDNA and microsatellites studies [15, 18-20, 24]. While the SNP assay was not explicitly designed to identify the eastern highlands walleye group, it nevertheless does distinguish walleye spawning in the Ohio River drainage (FOS and ROC) as genetically distinct, reflecting their historical isolation [21, 25]. However, finer-scale demarcations of walleye populations across the upper northern regions (Northwest Lake Plains, Great Lakes watershed, and North Atlantic coastal) are not recovered. Future work could develop SNP resources for characterizing fine-scale population structure through GBS sequencing of all five major walleye lineages or by using existing GBS data generated from other northern walleye populations [61, 62].

4.3 Identification of Southern and Northern Walleye Hybrids

Characterizing and detecting the genomic composition of hybrids is critical for studies of hybrid zone dynamics, inheritance of traits, and consequences of stocking and hybridization for evolution, fishery management and conservation [63]. Empirical data and simulations have demonstrated that 50 or more ancestry-informative markers are needed to accurately identify F_2 hybrids and advanced-generation backcross individuals [63, 64]. For instance, a recent bighead (*Hypophthalmichthys nobilis*) and silver (*H. molitrix*) carp study using 57 diagnostic SNPs successfully identified advanced-stage hybrids throughout their distribution in the Mississippi River Basin [6]. In this study, a SNP assay with 68 diagnostic markers was developed for rapid and accurate identification of genetic purity and classification of various (northern/southern) hybrid classes among walleye individuals. The precision of hybrid classification using the SNP panel was evaluated by three main aspects: reliability of genotypes, the accuracy of simulations, and the repeatability of assignments between GBS and MassARRAY data. Owing to the stringent criteria applied for assay development, I observed high concordance between GBS and MassARRAY genotypes (99.75%), and found only 9 discrepancies out of 7250 genotype comparisons among technical replicates (114 samples genotyped twice with MassARRAY), suggesting the high reliability and repeatability of the genotyping data. In both cases, the genotype discrepancy did not impact the assignment results. Secondly, simulation analyses using GBS and MassARRAY datasets showed that the SNP markers have enough discriminatory power to correctly identify up to third-generation hybrids (> 89% accuracy). Similar simulation results based on 96 species-specific SNPs were previously reported for two North Atlantic eel species (*Anguilla anguilla* and *A. rostrate*; [65]). In addition, I only found two discrepancies when I compared the hybrid assignment results generated from GBS (300 fixed SNPs) and the 68-SNP assay. The assignment discrepancy was potentially due to the lack of homozygous loci in the least genetically represented hybrid classes in MassARRAY data (i.e., B_x3S × B_x3S only possesses an average of 1.56% or 0.27 of 68 loci homozygous for northern walleye, [6]). In this study, because MassARRAY data failed to detect rare loci homozygous for northern walleye, two individuals (among 22 hybrids) identified as F_xS category from GBS data were mis-assigned to B_x2S with 68 MassARRAY markers.

Walleye from the Black Warrior River were mostly later-generation hybrids (92.7%, Table 7), which is unsurprising given the long history of non-native walleye stocking in this watershed. Between 1975 and 1985, thousands of northern walleye fingerlings sourced from Seneca Lake (Ohio) and Pymatuning Lake (Pennsylvania) were stocked into Tuscaloosa Reservoir (namely the North River population) and Sipsey River [26]. Previous mitochondrial DNA analyses found evidence of walleye hybridization along these waterways and hypothesized that the introduced fish could potentially pass over the Tuscaloosa dam or swim upstream in the Black Warrior River and hybridize with native southern walleye [26]. Given that the generation interval approximates the mean age of breeding individuals for populations with overlapping generations [66], it is expected that multiple generations of hybridization have occurred in the Black Warrior River.

4.4 Conservation Implication in Southern Walleye

Given the significant genetic distinctiveness of walleye in the Mobile River Basin, conserving these populations is critical for resource management and preserving biodiversity. Any efforts at conservation through stocking and genetic rescue for these populations should assess the genomic purity of donor walleye stocks [50]. The SNP panel and MassARRAY system offer a cost-effective and reliable tool for this purpose, with the 68-SNP assay already being implemented in ongoing stream survey and captive breeding programs. Meanwhile, careful monitoring of genetic diversity between donor and recipient walleye stocks should be conducted to stall the erosion of genetic diversity and enhance the long-term survival of southern walleye in the wild. This could be accomplished by designing additional polymorphic SNP panels from the *intermediate* SNP dataset.

Data analysis of historical demography suggests that the southern walleye population in Hatchet Creek has undergone a continuous decline in population size over ~800,000 generations. The low effective population size estimated from both demographic (Fig. 2) and genetic diversity (Table 3) analyses suggest a high risk of imminent local extinction, as genetic drift in small populations can have a great influence on genetic diversity and population fitness [67, 68]. The declining trend in southern walleye population size is also reflected by the low catch rate of wild walleye throughout the Alabama River systems (e.g., only 31 southern walleye were collected in a 2-year survey, [22]), and rare spawning events reported in Luxapallila Creek from the Tombigbee River [69]. Several natural and anthropogenic factors may be contributing to the low N_e estimate for the Hatchet Creek walleye population. First, walleye in their natural environment are characterized by low and unequal reproductive success and high mortality rate, all of which negatively affect the effective population size [70]. In addition, ecosystem and community changes, including the introduction of predators or competitors, habitat degradation, climate change, and altered hydrologic conditions, may influence rates of growth, survival, and recruitment in the southern walleye populations [71]. Lastly, fishery or angling exploitation has been proposed as a significant risk to walleye populations by negatively impacting recruitment variability, growth rate and age to maturity [72, 73]. Given the continued threat of habitat loss and climate change, complementary data related to life-history features (e.g., generation time, spawning success, survival rate) and population dynamics (e.g., census population size, exploitation rate) in southern walleye need to be collected to facilitate the conservation and management of this unique group. As a final point, it is essential to highlight that while stocking and genetic rescue processes [74] can help facilitate the conservation and management of southern walleye, they do not address issues of habitat loss and ecological degradation. A well-designed restoration strategy and strict regulation are necessary to help recover the ecosystem and guard against the extinction of southern walleye [75].

References

1. Dudgeon, D., et al., Freshwater biodiversity: importance, threats, status and conservation challenges. Biological reviews of the Cambridge Philosophical Society, 2006. **81**(2): p. 163-182.
2. Cochran‐Biederman, J.L., et al., Identifying correlates of success and failure of native freshwater fish reintroductions. Conservation Biology, 2015. **29**(1): p. 175-186.
3. Miller, L., Genetic guidelines for hatchery supplementation programs. Population Genetics Principles and Applications for Fisheries Science, 2003.
4. Seddon, P.J., D.P. Armstrong, and R.F. Maloney, Developing the science of reintroduction biology. Conservation Biology, 2007. **21**(2): p. 303-312.
5. Andrews, K.R., et al., Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 2016. **17**(2): p. 81.
6. Lamer, J.T., et al., Diagnostic SNPs reveal widespread introgressive hybridization between introduced bighead and silver carp in the Mississippi River Basin. Molecular Ecology, 2015. **24**(15): p. 3931-3943.
7. Pritchard, V.L., et al., Single nucleotide polymorphisms to discriminate different classes of hybrid between wild Atlantic salmon and aquaculture escapees. Evolutionary Applications, 2016. **9**(8): p. 1017-1031.
8. Thongda, W., et al., Species-diagnostic SNP markers for the black basses (*Micropterus spp.*): a new tool for black bass conservation and management. Conservation Genetics Resources, 2019: p. 1-10.

9. Thongda, W., et al., Development of SNP Panels as a New Tool to Assess the Genetic Diversity, Population Structure, and Parentage Analysis of the Eastern Oyster (*Crassostrea virginica*). Marine Biotechnology, 2018. **20**(3): p. 385-395.

10. Zhao, H., et al., SNP panel development for genetic management of wild and domesticated white bass (*Morone chrysops*). Animal Genetics, 2019. **50**(1): p. 92-96.

11. Barton, B.A., Biology, management, and culture of walleye and sauger. 2011: American Fisheries Society Bethesda, Maryland.

12. Regier, H.A., V.C. Applegate, and R.A. Ryder, The ecology and management of the walleye in western Lake Erie. Great Lakes Fishery Commission. Technical Report, 1969(15): p. I-101.

13. Hokanson, K.E., Temperature requirements of some percids and adaptations to the seasonal temperature cycle. Journal of the Fisheries Board of Canada, 1977. **34**(10): p. 1524-1550.

14. Billington, N. and P.D. Hebert, Mitochondrial DNA variation in Great Lakes walleye (*Stizostedion vitreum*) populations. Canadian Journal of Fisheries and Aquatic Sciences, 1988. **45**(4): p. 643-654.

15. Stepien, C.A. and J.E. Faber, Population genetic structure, phylogeography and spawning philopatry in walleye (*Stizostedion vitreum*) from mitochondrial DNA control region sequences. Molecular Ecology, 1998. **7**(12): p. 1757-1769.

16. Ward, R.D., N. Billington, and P.D. Hebert, Comparison of allozyme and mitochondrial DNA variation in populations of walleye, *Stizostedion vitreum.* Canadian Journal of Fisheries and Aquatic Sciences, 1989. **46**(12): p. 2074-2084.

17. Billington, N. and R.M. Strange, Mitochondrial DNA analysis confirms the existence of a genetically divergent walleye population in northeastern Mississippi. Transactions of the American Fisheries Society, 1995. **124**(5): p. 770-776.

18. Billington, N. Geographical distribution of mitochondrial DNA (mtDNA) variation in walleye, sauger, and yellow perch. in Annales Zoologici Fennici. 1996. JSTOR.

19. Stepien, C.A., et al., Signatures of vicariance, postglacial dispersal and spawning philopatry: population genetics of the walleye *Sander vitreus*. Molecular Ecology, 2009. **18**(16): p. 3411-3428.

20. Billington, N., R.J. Barrette, and P.D. Hebert, Management implications of mitochondrial DNA variation in walleye stocks. North American Journal of Fisheries Management, 1992. **12**(2): p. 276-284.

21. White, M.M., J.E. Faber, and K.J. Zipfel, Genetic identity of walleye in the Cumberland River. The American Midland Naturalist, 2012. **167**(2): p. 373-384.

22. Billington, N., R.M. Strange, and M.J. Maceina. Mitochondrial-DNA confirmation of southern walleye in the Mobile Basin, Alabama. in Proceedings of the Annual Conference Southeastern Association of Fish and Wildlife Agencies. 1997.

23. Murphy, B.R. Evidence for a genetically unique walleye population in the upper Tombigbee River system of northeastern Mississippi. in Southeastern Fishes Council Proceedings. 1990.

24. Haponski, A.E. and C.A. Stepien, A population genetic window into the past and future of the walleye *Sander vitreus*: relation to historic walleye and the extinct “blue pike” S. v.“glaucus”. BMC Evolutionary Biology, 2014. **14**(1): p. 133.

25. Palmer, G.C., et al., Genetic distinct walleye stocks in Claytor Lake and the upper New River, Virginia. P Southeast Fish Wild Agencies, 2006. **60**: p. 125-131.

26. Billington, N. and M.J. Maceina, Genetic and population characteristics of walleyes in the Mobile drainage of Alabama. Transactions of the American Fisheries Society, 1997. **126**(5): p. 804-814.

27. Zimin, A.V., et al., The MaSuRCA genome assembler. Bioinformatics, 2013. **29**(21): p. 2669-2677.

28. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009. **25**(14): p. 1754-1760.

29. Catchen, J., et al., Stacks: an analysis tool set for population genomics. Molecular Ecology, 2013. **22**(11): p. 3124-3140.

30. Li, H., et al., The sequence alignment/map format and SAMtools. Bioinformatics, 2009. **25**(16): p. 2078-2079.

31. Danecek, P., et al., The variant call format and VCFtools. Bioinformatics, 2011. **27**(15): p. 2156-2158.

32. Zheng, X., et al., A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics, 2012. **28**(24): p. 3326-3328.

33. Li, C., et al., SNP discovery in wild and domesticated populations of blue catfish, *Ictalurus furcatus*, using genotyping-by-sequencing and subsequent SNP validation. Molecular Ecology Resources, 2014. **14**(6): p. 1261-1270.

34. Peakall, R. and P.E. Smouse, GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. Bioinformatics (Oxford, England), 2012. **28**(19): p. 2537-2539.

35. Foll, M. and O. Gaggiotti, A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics, 2008. **180**(2): p. 977-993.

36. Zhao, H., et al., SNP marker panels for parentage assignment and traceability in the Florida bass (*Micropterus floridanus*). Aquaculture, 2018. **485**: p. 30-38.

37. Pompanon, F., et al., Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics, 2005. **6**(11): p. 847.

38. Pritchard, J.K., M. Stephens, and P. Donnelly, Inference of population structure using multilocus genotype data. Genetics, 2000. **155**(2): p. 945-959.

39. Kopelman, N.M., et al., Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources, 2015. **15**(5): p. 1179-1191.

40. Hudson, R.R., M. Slatkin, and W.P. Maddison, Estimation of levels of gene flow from DNA sequence data. Genetics, 1992. **132**(2): p. 583-589.

41. Patterson, N., A.L. Price, and D. Reich, Population structure and eigenanalysis. PLoS Genetics, 2006. **2**(12): p. e190.

42. Bhatia, G., et al., Estimating and interpreting FST: the impact of rare variants. Genome Research, 2013. **23**(9): p. 1514-1521.

43. Jombart, T. and I. Ahmed, adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics, 2011. **27**(21): p. 3070-3071.

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

45. Do, C., et al., NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular Ecology Resources, 2014. **14**(1): p. 209-214.

46. Anderson, E.C. and E.A. Thompson, A model-based method for identifying species hybrids using multilocus genetic data. Genetics, 2002. **160**(3): p. 1217-1229.

47. Elliott, L. and M.A. Russello, SNP panels for differentiating advanced-generation hybrid classes in recently diverged stocks: A sensitivity analysis to inform monitoring of sockeye salmon re-stocking programs. Fisheries Research, 2018. **208**: p. 339-345.

48. Nielsen, E.E., L.A. Bach, and P. Kotlicki, HYBRIDLAB (version 1.0): a program for generating simulated hybrids from population samples. Molecular Ecology Notes, 2006. **6**(4): p. 971-973.

49. Excoffier, L., et al., Robust demographic inference from genomic and SNP data. PLoS Genetics, 2013. **9**(10): p. e1003905.

50. Chattopadhyay, B., et al., Conservation genomics in the fight to help the recovery of the critically endangered Siamese crocodile *Crocodylus siamensis*. Molecular Ecology, 2019.

51. Nachman, M.W. and S.L. Crowell, Estimate of the mutation rate per nucleotide in humans. Genetics, 2000. **156**(1): p. 297-304.

52. Malinsky, M., et al., Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nature Ecology & Evolution, 2018. **2**(12): p. 1940.

53. Feng, C., et al., Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. Elife, 2017. **6**: p. e23907.

54. Evanno, G., S. Regnaut, and J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 2005. **14**(8): p. 2611-2620.

55. Hare, M.P., et al., Understanding and estimating effective population size for practical application in marine species management. Conservation Biology, 2011. **25**(3): p. 438-449.

56. Ogden, R., et al., Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. Molecular Ecology, 2013. **22**(11): p. 3112-3123.

57. Harrison, R.G. and E.L. Larson, Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Molecular Ecology, 2016. **25**(11): p. 2454-2466.

58. Narum, S.R., et al., Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology, 2013. **22**(11): p. 2841-2847.

59. Petit, R.J., et al., Glacial refugia: hotspots but not melting pots of genetic diversity. Science, 2003. **300**(5625): p. 1563-1565.

60. Stepien, C.A., O.J. Sepulveda-Villet, and A.E. Haponski, Comparative genetic diversity, population structure, and adaptations of walleye and yellow perch across North America, in Biology and Culture of Percid Fishes. 2015, Springer. p. 643-689.

61. Allen, B.E., et al., Loss of SNP genetic diversity following population collapse in a recreational walleye (*Sander vitreus*) fishery. Canadian Journal of Fisheries and Aquatic Sciences, 2017. **75**(10): p. 1644-1651.

62. Chen, K.-Y., Lake Erie walleye population structure and stock discrimination methods. 2016, The Ohio State University.

63. Fitzpatrick, B.M., Estimating ancestry and heterozygosity of hybrids using molecular markers. BMC Evolutionary Biology, 2012. **12**(1): p. 131.

64. Malde, K., et al., Whole genome resequencing reveals diagnostic markers for investigating global migration and hybridization between minke whale species. BMC Genomics, 2017. **18**(1): p. 76.

65. Pujolar, J.M., et al., Assessing patterns of hybridization between North Atlantic eels using diagnostic single-nucleotide polymorphisms. Heredity (Edinb), 2014. **112**(6): p. 627.

66. Hill, W.G., A note on effective population size with overlapping generations. Genetics, 1979. **92**(1): p. 317-322.

67. Franckowiak, R., et al., Temporal effective size estimates of a managed walleye *Sander vitreus* population and implications for genetic-based management. Journal of Fish Biology, 2009. **74**(5): p. 1086-1103.

68. Willi, Y., J. Van Buskirk, and A.A. Hoffmann, Limits to the adaptive potential of small populations. Annual Review of Ecology, Evolution, and Systematics., 2006. **37**: p. 433-458.

69. Schramm Jr, H.L., J. Hart, and L.A. Hanson, Status and reproduction of Gulf Coast strain walleye in a Tombigbee River tributary. Southeastern Naturalist, 2004: p. 745-757.

70. Ivan, L.N., et al., Density, production, and survival of walleye (*Sander vitreus*) eggs in the Muskegon River, Michigan. Journal of Great Lakes Research, 2010. **36**(2): p. 328-337.

71. Nate, N.A., et al., Population and community dynamics of walleye. Biology, management, and culture of walleye and sauger. American Fisheries Society, Bethesda, Maryland, 2011: p. 321-374.

72. Baccante, D.A. and P.J. Colby. Harvest, density and reproductive characteristics of North American walleye populations. in Annales Zoologici Fennici. 1996. JSTOR.

73. Spangler, G., et al., Responses of percids to exploitation. Journal of the Fisheries Board of Canada, 1977. **34**(10): p. 1983-1988.

74. Whiteley, A.R., et al., Genetic rescue to the rescue. Trends in Ecology & Evolution, 2015. **30**(1): p. 42-49.

75. Peterson, G.D., G.S. Cumming, and S.R. Carpenter, Scenario planning: a tool for conservation in an uncertain world. Conservation Biology, 2003. **17**(2): p. 358-366.