MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12272

SNP discovery in wild and domesticated populations of blue catfish, *Ictalurus furcatus*, using genotyping-by-sequencing and subsequent SNP validation

CHAO LI,* GEOFF WALDBIESER,† BRIAN BOSWORTH,† BENJAMIN H. BECK,‡ WILAWAN THONGDA* and ERIC PEATMAN*

*School of Fisheries, Aquaculture and Aquatic Sciences, Auburn University, Auburn, AL 36849, USA, †Warmwater Aquaculture Research Unit, USDA-ARS, 141 Experiment Station Road, Stoneville, MS 38776, USA, ‡United States Department of Agriculture, Agricultural Research Service, Stuttgart National Aquaculture Research Center, Stuttgart, AR 72160, USA

Abstract

Blue catfish, *Ictalurus furcatus*, are valued in the United States as a trophy fishery for their capacity to reach large sizes, sometimes exceeding 45 kg. Additionally, blue catfish × channel catfish (*I. punctatus*) hybrid food fish production has recently increased the demand for blue catfish broodstock. However, there has been little study of the genetic impacts and interaction of farmed, introduced and stocked populations of blue catfish. We utilized genotyping-by-sequencing (GBS) to capture and genotype SNP markers on 190 individuals from five wild and domesticated populations (Mississippi River, Missouri, D&B, Rio Grande and Texas). Stringent filtering of SNP-calling parameters resulted in 4275 SNP loci represented across all five populations. Population genetics and structure analyses revealed potential shared ancestry and admixture between populations. We utilized the Sequenom MassARRAY to validate two multiplex panels of SNPs selected from the GBS data. Selection criteria included SNPs shared between populations, SNPs specific to populations, number of reads per individual and number of individuals genotyped by GBS. Putative SNPs were validated in the discovery population and in two additional populations not used in the GBS analysis. A total of 64 SNPs were genotyped successfully in 191 individuals from nine populations. Our results should guide the development of highly informative, flexible genotyping multiplexes for blue catfish from the larger GBS SNP set as well as provide an example of a rapid, low-cost approach to generate and genotype informative marker loci in aquatic species with minimal previous genetic information.

Keywords: blue catfish, genotyping-by-sequencing, Sequenom, SNP

Received 20 March 2014; revision received 28 April 2014; accepted 28 April 2014

Introduction

Blue catfish (*Ictalurus furcatus*) are the largest catfish species in North America and one of the largest endemic freshwater species, capable of reaching sizes exceeding 45 kg over a lifespan of >20 years (Graham 1999). They are native to the Mississippi, Missouri and Ohio River basins of the central and southern United States (Glodek 1980), where they are valued for their trophy recreational and commercial value (Arterburn *et al.* 2002; Bartram 2010). Concerted efforts to stock blue catfish have been made in states such as Texas, where >10 000 000 fingerlings have been released (Bartram 2010), while they have been introduced widely outside their native range by anglers (Schloesser *et al.* 2011). Additionally, farm culture of a blue catfish × channel catfish (*Ictalurus punctatus*) artificial hybrid (Giudice 1966; Bosworth & Waldbieser 2014) has increased dramatically in the southeast United States in the past decade, necessitating the development and maintenance of blue catfish broodstock lines and strains.

A small number of studies have begun to examine the ecological roles of blue catfish on fish assemblages in both native and introduced ranges. However, these studies are complicated by the relatively shallow depth of knowledge and resources available for the species (Graham 1999; Eggleton & Schramm 2004; Bartram 2010; Schloesser *et al.* 2011). One of the factors limiting large-scale analyses of blue catfish populations is the lack of molecular marker resources. An earlier study utilized AFLP markers (Liu *et al.* 1998), and a handful of non-peer-reviewed studies have adapted small numbers of microsatellites from channel catfish (e.g. Grist 2002). More

Correspondence: Eric Peatman, Fax: +001-334-844-4694; E-mail: peatmer@auburn.edu

recently, however, focus has turned to the utility of SNP markers, valued for their even genome-wide distribution, abundance and low genotyping error rate for high-throughput analyses (Slate *et al.* 2009). Advances in sequencing technology and associated dramatic declines in sequencing cost have greatly facilitated large-scale discovery of SNP markers (Ekblom & Galindo 2011; Vandepitte *et al.* 2013). However, efforts in most nonmodel organisms have been, until recently, limited to transcriptome sequencing of pooled samples to reduce genome complexity and costs associated with library preparation of individual samples. In this vein, a set of putative SNPs were identified from a single pooled RNA sample of 19 blue catfish from two cultured strains (Liu *et al.* 2011). However, given the low stringency call parameters (minimum coverage read depth of four and the minimum variant frequency of two), restriction to the genic region, lack of individual genotypes, lack of validation and limited source diversity, the utility of this previous study to population genetics and genomics analyses was limited.

Several methods that reduce the complexity of the genome while avoiding ascertainment bias (Morin *et al.* 2004) stemming from screening too few individuals for SNP markers have emerged in the last several years. These methods seek to achieve a balance between sequence space, number of individual samples and depth of sequence at a particular locus (Krück *et al.* 2013). Reduced-representation library (RRL) sequencing approaches have proven successful in allowing cost-effective SNP discovery and allele frequency estimation across multiple individuals (Van Tassell *et al.* 2008). A derivation of this approach, restriction-site-associated DNA sequencing (RAD-seq), has particularly gained popularity in nonmodel organisms (Baird *et al.* 2008; Gompert *et al.* 2010; Krück *et al.* 2013; Vandepitte *et al.* 2013). Numerous variants of this technique or similar approaches have also been developed recently (e.g. PE-RAD, 2b-RAD, ezRAD). Another RRL-based approach is genotyping-by-sequencing (GBS) which, like RAD-seq, reduces genome complexity through restriction digest, but which offers a simplified and more cost-effective library preparation protocol (Davey *et al.* 2011; Elshire *et al.* 2011; De Donato *et al.* 2013).

Here, we utilized GBS for genome-wide SNP identification and genotyping of 190 blue catfish individuals from five populations (both wild and domesticated). We analysed the utility of a stringently filtered SNP set for population-level differentiation and identified allelic imbalances in each population. Given our desire to utilize sets of SNP markers in the near future for large-scale population genetics studies, we examined how individual- and read-level coverage at a particular locus impacted rates of validation in multiplex panels designed for the Sequenom MassARRAY platform. The SNP resources reported here should be of high utility in future studies examining genetic diversity and population differentiation, genome signatures of domestication and selection, genetic origin of introduced or farmed fish, and the impact of escapees from aquaculture facilities (Van Bers *et al.* 2012).

Materials and methods

Sample collection and DNA extraction

A total of 190 blue catfish individual blood samples were collected from five strains/populations for GBS analysis (Fig. 1). Of note, the wild origin of commercial strains of blue catfish is not always clear, nor is their breeding history since domestication. Domesticated strain origins are described according to Dunham and Smitherman (1984). The five populations were as follows: (i) 48 D&B individuals from a commercial farm in Arkansas (strain originating from D&B Fish Farm in Crockett, TX, and generated by crossing females from the Trinity River, TX, with males from the Mississippi River in 1963); (ii) 47 Rio Grande individuals from a breeding population at the Stuttgart National Aquaculture Research Center, Stuttgart, AR (strain originating from the Rio Grande River, TX, in the early 1970s); (iii) 32 Missouri individuals from a commercial farm in Missouri (wild source unknown); (iv) 32 individuals from the Jasper State Fish Hatchery, Jasper, TX (broodstock originally captured from the Trinity River, TX); (v) 31 Mississippi River individuals from wild spawns collected from the river near Vicksburg, MS. For subsequent validation analysis, we

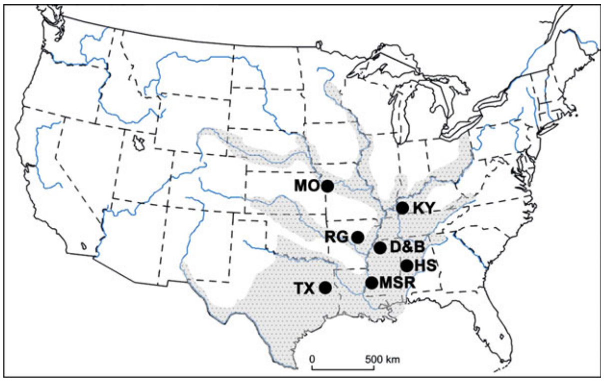

Fig. 1 Blue catfish sample origins for this study overlaid on a map of the historical, native range of blue catfish, as modified from Dunham & Smitherman (1984). MO is Missouri, KY is Kentucky Lake, KY, RG is Rio Grande strain, D&B is D&B strain, HS is Harvest Select strain, MSR is Mississippi River, and TX is Texas hatchery samples. Note that sample origin may not correspond with ancestral origin. Greater details of sample origins and numbers are provided in the Materials and Methods.

additionally genotyped 24 individuals caught by a commercial fisherman in Kentucky Lake, KY, and 23 putative D&B strain samples sourced from Harvest Select Catfish (Uniontown, AL).

Genomic DNA was extracted from blue catfish blood samples 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.

Genotyping-by-sequencing sample preparation and sequencing

Library preparation and sequencing protocols followed those of Elshire *et al.* (2011), 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 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 *Eco*T22I (New England Biolabs, Ipswitch, MA) using the manufacturer's recommended conditions. *EcoT22I* is a 6-bp cutting enzyme previously shown to work well for populations of unknown structure and highly heterozygous materials (Chen *et al.* 2013). 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 and purified using a commercial kit (QIAquick PCR Purification Kit), and then restriction fragments with ligated adapters were amplified to generate sequencing libraries. Libraries were purified as above and fragment sizes evaluated on an ExperionH 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.

The GBS analysis pipeline [TASSEL version: 3.0.147; (Glaubitz *et al.* 2014)] was run on all the samples (default settings). Briefly, sequences were trimmed of ambiguous nucleotides and barcodes first. Potentially 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 blue catfish contigs by Burrows-Wheeler Aligner (BWA) separately (Li & Durbin 2009), 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 SNP detection, minimum minor allele frequency was set to 0.01 (overall), minimum minor allele count was set to 10 (per population), and minimum locus coverage was set to 0.1 (overall). The minor allele frequency of 0.01 allowed for a single minor allele call in a SNP with 100 reads, dictating that a minor allele count would have to be ≥10, eliminated potential sequencing errors and meant that at least 1000 reads were needed for a filtered SNP. Additionally, in order to be included in subsequent analyses, a SNP had to be biallelic and genotyped in ≥70% of individuals within each population. Finally, to avoid potentially paralogous loci not identified by BLAST alignments, we only utilized loci with an observed heterozygosity (H_o) of ≤0.6 in each population for downstream analyses. To identify gene-associated SNPs, the FGENESH program was used to predict protein-coding sequences of each genomic contig with SNPs (Salamov & Solovyev 2000). The predicted protein-coding sequences were then annotated using BLASTp searches against the Uniprot database. The SNPs in the coding region and with a significant gene hit were considered as gene-associated SNPs.

Population-level analyses of genotype data

To reduce the number of closely linked markers, only the SNP with the highest number of scored individuals was selected from genomic contigs in which multiple SNP loci were mapped. The filtered SNP set meeting these parameters was used to perform downstream population genetics tests. STRUCTURE version 2.3.4 was used to estimate the number of populations (K) in all the sequenced populations (Pritchard *et al.* 2000). Briefly, a Markov chain Monte Carlo simulation was run for 10^5 iterations following a burn-in period of 10^5 iterations for K = 1–5 using the correlated allele frequencies model and assumed admixture with 15 replications for each K. The STRUCTURE-generated results were imported to Structure Harvester software to determine the best K (Earl 2012). The generated Q-matrix under the best K from STRUCTURE was exported to DISTRUCT version 1.1 to generate bar plots to depict classifications with the highest probability (Rosenberg 2004). Phylogenetic relationships were constructed using POPULATIONS version 1.2.32 with the neighbour-joining (NJ) method and Nei's standard genetic distance *Ds* using the grouped population option (Langella 1999). ARLEQUIN version 3.5.1.3 was used to calculate genetic diversity among populations by calculating the observed (H_o) and expected heterozygosity (H_e), and the within-population fixation index (F_{IS}) (Excoffier & Lischer 2010). Hardy–Weinberg equilibrium (HWE; exact test using a Markov chain with chain length = 1 000 000 and dememorization steps = 100 000) was tested in each population. An analysis of molecular variance (AMOVA) was performed with number of permutations = 10 000 to calculate the distribution of variance. Pairwise F_{ST} values with Bonferroni correction were assessed using GENEPOP version 4.2 (Rousset 2008).

Validation of GBS SNPs

The Sequenom MassARRAY platform (Sequenom, San Diego, CA, USA) was employed to validate a subset of SNPs identified and genotyped by GBS. A total of 191 individuals were genotyped using two multiplex panels: 23 D&B, 26 Rio Grande, 32 Missouri, 32 Texas, 31 Mississippi River, 24 Harvest Select and 23 Kentucky Lake individuals. Sample sources were as described above. Samples utilized for the GBS sequencing as well as new samples from those populations (*n* = 7) were used. Sequenom assays were designed using the MassARRAY Assay Design Software with the goal of maximizing multiplexing of 40 SNPs per well. Only SNPs with at least a 100-bp flanking region on either side of the polymorphism site were selected for the assay design. 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 Sequenom System Typer 4.0 analysis software. This software uses a three-parameter 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 the relative significance. Noncalls also were noted (e.g. low probability, bad spectrum).

Results and discussion

Genotyping-by-sequencing – marker discovery and genotyping

The origins and relationships of wild and domesticated blue catfish strains and populations in the United States are currently far from clear (Graham 1999; Grist 2002). A two-way exchange of genetic material between wild and domesticated stocks has occurred as farmed and hatchery-reared fish were sourced from the wild and then re-introduced to varying extents through state stocking programmes, farm escapes and transfer to new rivers and reservoirs by anglers. Farmed blue catfish strains were generated from wild fish in the 1960s and 1970s and have been widely distributed across the southeast with the growth of catfish farming. These farmed strains have likely diverged substantially from source populations due to domestication pressure, trait selection and admixture with other strain populations. Given these realities, our aim here was to generate SNP marker resources with a high utility for population management of wild, domesticated and hatchery-reared blue catfish populations. We utilized 190 samples from state hatchery, farmed and wild populations for genotyping-by-sequencing analysis (Fig. 1). Genotyping-by-sequencing has been utilized in plants (Kopecký & Studer 2013) and cattle (De Donato *et al.* 2013), but this represents, to our knowledge, the first application of the approach to aquatic organisms. A total of 348 150 896 high-quality reads (428 128 488 raw reads) were generated from two Illumina HiSeq lanes. The TASSEL-GBS pipeline (Glaubitz *et al.* 2014) clustered reads into 1 488 960 locus-specific tags with mean tag coverage of 23 reads/individual. During the BWA alignment to the blue catfish draft genome (available upon request), 469 012 tags were aligned to unique positions (31.50%), 626 151 tags were aligned to multiple positions (42.05%), and 393 797 tags could not be aligned (26.45%). Given the incomplete nature of the blue catfish assembly used (88 864 contigs with 13 271-bp average contig size), and the short size of the tags, this result was largely expected. A larger mapping population of tags may be able to be gleaned from these data in the future as the blue catfish assembly improves.

We carried forward the uniquely mapped tags for SNP detection. A total of 21 145 unfiltered SNPs were identified from the TASSEL package. A final set of 4275 SNPs were retained after passing our more stringent parameters (Table S1, Supporting information). To call a SNP within an individual, a minimum minor allele frequency (MAF) of 0.01 and a minimum number minor allele read count (MAC) of 10 criteria had to be met. Additionally, a SNP locus had to be present and scorable in 70% of the individuals within each population. We finally filtered out potentially paralogous loci with H_o values of >0.6. Of these filtered SNPs, a portion was polymorphic in only one population, including 1137 in D&B, 28 in Rio Grande, 169 in Missouri, 163 in Texas and 153 in Mississippi River (Fig. 1, Table S2, Supporting information). The higher number of D&B-specific polymorphic SNPs may reflect the founding cross of Mississippi River and Trinity River, TX catfish as well as continued admixture in the domesticated strain. Of note, at many of the examined loci, only a small number of D&B individuals were polymorphic, such that setting a parameter requiring >10% of individuals be polymorphic at a given locus reduced this number to 528 (data not shown). While no fixed population-specific SNPs were identified, the identified SNPs with skewed allele use ratios, if used together, could be useful in tracing fish back to populations of origin.

A total of 791 SNPs were polymorphic in all five populations, with smaller numbers polymorphic in two to four populations (Fig. 2, Table S2, Supporting information). A slight reduction in SNP numbers was observed when we screened for SNPs with 100-bp flanking regions on each side of the SNP locus (4012). Using FGENESH, we identified 1020 SNPs in a predicted protein-coding region (Table S3, Supporting information).

Genetic diversity and variation in blue catfish

The percentage of polymorphic loci (P_o) and observed and expected heterozygosity (H_o and H_e respectively) for all 4275 loci was estimated for each population using ARLEQUIN. Reflecting the higher number of loci polymorphic in D&B alone (Table 1), D&B individuals were polymorphic at the highest percentage of loci (58.97%), with other populations polymorphic at between 39% and 52% of loci. Average observed heterozygosity (H_o) among all populations ranged from 0.09 to 0.11 with a mean of 0.09 (Table 1). The highest observed heterozygosity was observed in Mississippi River samples at 0.11. There were no significant differences between observed and expected heterozygosity. After testing each loci for deviation from Hardy–Weinberg equilibrium (HWE), an average of 13.69% of loci was found to depart from HWE. We also calculated inbreeding coefficients (F_{IS}) for each of the populations. The highest F_{IS} value was calculated from D&B (0.30, significantly different than 0, P = 0.01), while the lowest value was observed from Missouri blue catfish (0.00). Based on the analysed samples, our results suggested highest, and moderate, levels of inbreeding in the long-domesticated D&B line, with no indication of inbreeding in other populations. Although this study represents the first population genetics assessment in Ictalurid catfish with SNP markers, smaller studies of diversity have been conducted using isozymes, AFLP or microsatellite markers in channel catfish (Hallerman *et al.* 1986; Waldbieser & Bosworth 1997; Tan *et al.* 1999; Mickett *et al.* 2003; Perales-Flores *et al.* 2007). Inbreeding levels here were higher in the domesticated D&B population, but lower in all other populations, when compared with levels reported in farmed populations in Mexico, from 0.17 to 0.27 (Perales-Flores *et al.* 2007; Parra-Bracamonte *et al.* 2011), with levels of heterozygosity and % polymorphic loci generally mirroring previous studies in catfish. A parallel study using historical catfish microsatellite loci alongside larger SNP panels may be needed in the future to aid in direct comparison between these studies.

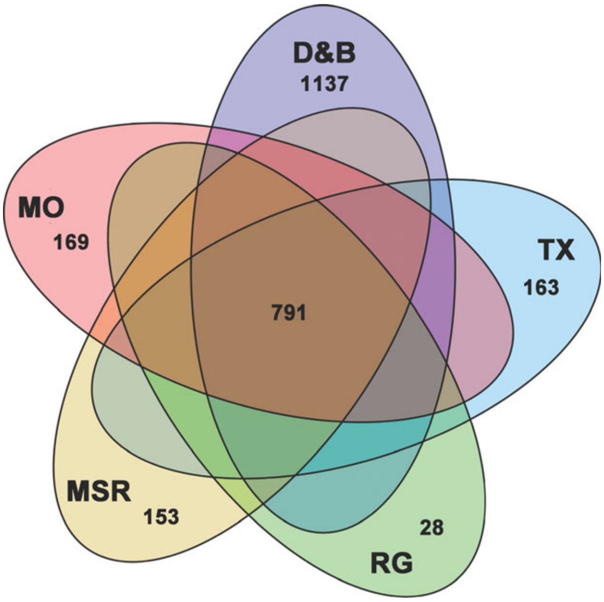

Fig. 2 Venn diagram of shared and population-specific SNPs identified from genotyping-by-sequencing (GBS) in blue catfish following filtering. Additional details are available in Table S2 (Supporting information).

Table 1 Genetic diversity parameters of 5 blue catfish populations based on 4275 SNP loci. P_o indicates percentage of polymorphic loci. H_o indicates average observed heterozygosity. H_e indicates average expected heterozygosity

Group	P_o %	H_o	H_e	% Loci Departure from HWE	F_{IS}
D&B	58.97	0.09	0.15	16.33	0.30*
Missouri	39.23	0.09	0.10	13.93	0.01
Mississippi River	45.02	0.11	0.12	12.51	0.00
Rio Grande	41.59	0.09	0.10	13.83	0.01
Texas	52.79	0.09	0.05	11.84	0.09
Mean	47.52	0.09	0.10	13.69	0.08
SD	0.07	0.01	0.03	0.02	0.12

*Significantly different from zero, based on 10 000 permutations.

To examine variance in allele frequencies among populations and their degree of differentiation, we measured the fixation index (F_{ST}) by GENEPOP. For this analysis, we used only a single SNP per genomic region (contig) to avoid the effects of linked data, resulting in 2214 SNPs (Table S4, Supporting information). The global F_{ST} across all the five populations was 0.18, indicating a moderate level of differentiation between populations. Pairwise F_{ST} values among all the five populations are listed in Table 2. The highest degree of differentiation was observed between the Rio Grande and Missouri

Table 2 Population pairwise F_{ST}s among the five blue catfish populations

	D&B	Missouri	Mississippi River	Rio Grande	Texas
D&B	0				
Missouri	0.23	0			
Mississippi River	0.16	0.24	0		
Rio Grande	0.17	0.28	0.19	0	
Texas	0.11	0.21	0.11	0.07	0

populations (F_{ST} = 0.28). Higher degrees of differentiation were also observed between Missouri and D&B (F_{ST} = 0.23), Missouri and Mississippi River (F_{ST} = 0.24) and Missouri and Texas (0.21), indicating the distinctness of the Missouri samples relative to other sampled blue catfish (Table 2). On the other hand, Texas catfish showed weak to moderate levels of population structuring with D&B (F_{ST} = 0.11), Mississippi River (F_{ST} = 0.11) and Rio Grande (F_{ST} = 0.07), indicating likely shared ancestry and admixture among these lines/populations. No previous study of blue catfish population structure has been conducted, but context and comparison can again be gained from previous studies of wild and farmed channel catfish. A study examining differentiation between farm stocks of channel catfish in one region of Mexico showed low levels of 0.04, reflecting the farming practice of sharing broodstock to maintain diversity (Perales-Flores *et al.* 2007). A similar study in a single, large Mexican hatchery observed an F_{ST} = 0.08 between temporally divided populations (Parra-Bracamonte *et al.* 2011). Similar values of F_{ST} to those obtained here have been reported using AFLP and isozyme analyses in domesticated channel catfish populations. Hallerman *et al.* (1986) found an F_{ST} of 0.24 among early channel catfish research lines, while Mickett *et al.* (2003) utilizing a larger number of AFLP loci calculated a global F_{ST} of 0.18 among 16 farmed Alabama populations. A study examining the genetic impact of domestic channel catfish on wild channel catfish populations reported a global F_{ST} value of 0.36, without reporting pairwise comparisons or partitioning values between wild and domesticated samples (Simmons *et al.* 2006). Given the known differences between F_{ST} values calculated with different marker types (Holsinger & Weir 2009), the likelihood of temporal shifts in population diversity, and limited historical records, further blue catfish sampling will be needed in the future to place our results in a more solid context.

Blue Catfish Population Structure Analysis

Using the same 2214 SNPs, a Bayesian model implemented in STRUCTURE software was used to examine the relationships between the five populations and assign individuals to clusters. A hypothetical population K value from 1 to 6 was tested. After passing the results to Structure Harvester program, the most probable clustering of the data was K = 3 (Fig. S1, Supporting information). The clusters identified were D&B; Missouri and Mississippi River; Rio Grande and Texas (Fig. 3). The phylogenetic relationships between populations were examined by the neighbour-joining (NJ) method in the POPULATIONS program and again indicated that the analysed populations belonged to three major clusters. The tree clustered the Missouri and the Mississippi River individuals (albeit with only moderate boot-strap support) and indicated a relationship between the Rio Grande and Texas individuals, while the relationship with the D&B population was again uncertain (Fig. 4). The higher level of admixture in D&B blue catfish relative to the other populations, as previously seen in high numbers of private alleles and polymorphic loci, was also apparent in the graphical representation of the STRUCTURE plot. The tangled, and usually undocumented, history of strain formation, stocking releases and fish transfers in blue catfish, combined with the lack of any previous survey of blue catfish genetic population structure, makes it hazardous to draw large-scale conclusions from the relationships revealed here.

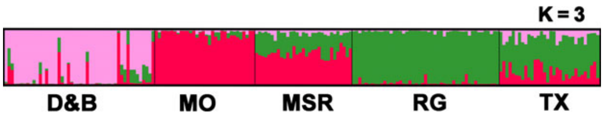

Fig. 3 Population structure of the five blue catfish GBS populations as assessed by STRUCTURE. Bar plot was generated by DISTRUCT. K (K = 3) indicates the number of clusters that maximized the probability of the model. Each individual is shown as a vertical bar.

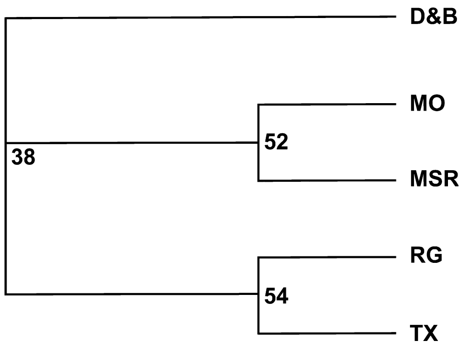

Fig. 4 A neighbour-joining tree constructed using the Nei's genetic distances calculated from pairwise comparisons of the genotypes (2214 SNPs) between the five blue catfish populations, generated by POPULATIONS version 1.2.32. Numbers at nodes are bootstrapping values with 10 000 permutations.

Nevertheless, our results demonstrate the utility of the SNP resources described here for tracing and differentiating blue catfish populations going forward.

Validation of GBS SNPs by Sequenom MassARRAY

While GBS represents a cost-effective tool for marker discovery and genotyping, regular field application of SNP markers in population genetic analysis of blue catfish will require the use of smaller, flexible multiplex panels for genotyping. Both Sequenom MassARRAY and Fluidigm-based multiplex genotyping solutions are increasingly utilized for conservation genetics and molecular ecology studies (Campbell & Narum 2011; Freamo *et al.* 2011; Pritchard *et al.* 2012; Krück *et al.* 2013). Here, we aimed to confirm GBS genotypes for SNPs selected across a range of read and individual coverages and with different patterns of polymorphism among populations. We additionally wished to confirm the utility of GBS SNPs in blue catfish from populations outside the GBS data set. We designed two Sequenom multiplex assays, each with 34 SNPs, for a total of 68 interrogated loci.

SNPs in the first assay were selected prior to filtering SNPs based on the parameters described above (Table S1, Supporting information). We selected SNPs in the first assay across a wide range of read coverages to understand the impact of GBS read coverage on SNP genotyping accuracy. SNPs were selected with total read coverage from 71 to 12 163 reads (Table S5, Supporting information). We also examined both population-specific and shared markers. Multiplex Assay 1 included four D&B-specific SNPs, five Missouri-specific SNPs, six Mississippi River-specific SNPs, nine Rio Grande-specific SNPs, three Texas-specific SNPs and seven shared (polymorphic in all populations) SNPs. Of the 34 attempted SNPs, only 17 were successfully amplified and shared the same allelic pattern as that predicted by the GBS results (i.e. validated). However, when total read count was taken into account, it was apparent that low read coverage explained most failing SNPs. While 81% of SNPs with >500 total reads were validated, only 7.6% of SNPs with <500 reads were validated (Table 3). In most cases, SNPs failing validation showed alleles predicted to occur only in a single population based on GBS in one or more additional populations. These results emphasized the need for higher locus read coverage to consistently capture minor alleles and the low reliability of GBS SNPs with <500 reads for downstream analyses. Accordingly, our SNP selection criteria resulted in 4275 SNPs with total read counts >500 reads in all cases (average total read count of 6205).

SNPs for Multiplex Assay 2 were those with >500 reads. However, here we also included nine SNPs with $H_o > 0.6$ in at least one population to test whether loci with higher than expected heterozygosity reflected loci under strong selection or potential paralogous sequences to be avoided. In this panel, we included seven D&B-specific SNPs, eight Missouri-specific SNPs, four Mississippi River-specific SNPs, four Rio Grande-specific SNPs, three Texas-specific-SNPs and eight shared (polymorphic in all populations) SNPs. All but one of the SNPs could be successfully amplified. Among the 33 amplified SNPs, 28 were validated to have the same allelic pattern as that observed by GBS (82% amplified and validated; Table S5, Supporting information; Table 3). We also ran both multiplexes on samples from two additional populations, one domestic (Harvest Select) and one wild (Kentucky Lake). All successfully amplified SNPs in the 5 GBS populations also amplified in the additional populations and, in most cases, were polymorphic in the

Table 3 Validation summary of blue catfish GBS SNPs tested in two multiplexes on the Sequenom MassARRAY platform. 'Validated' designates SNPs which amplified and in which the same allelic patterns were observed by Sequenom as were indicated in the GBS data set. Individual call rate was calculated based on the 64 successfully amplified SNPs

	Designed	Successfully Amplified	Validated	Successful validation rate[a]/ Individual call rate[b] %
Multiplex Assay 1 (<500 reads)	13	13	1	7.69[a]
Multiplex Assay 1 (>500 reads)	21	18	17	80.95[a]
Multiplex Assay 2 (>500 reads)	34	33	28	82.35[a]
D&B	11	10	7	93.72[b]
Missouri	13	13	12	93.11[b]
Mississippi River	10	10	9	94.45[b]
Rio Grande	13	12	4	93.81[b]
Texas	6	6	3	85.78[b]
Shared	15	13	10	93.68[b]
Total	68	64	45	92.97[b]

[a]represent the successful validation rate.
[b]represent the individual call rate.

'shared' group of markers (Table S6, Supporting information). Four of the five SNPs failing to validate came from the nine SNPs with $H_o > 0.6$, indicating that while some of these loci may represent true SNPs, they would be better avoided in future SNP panels.

Our primary purpose in the validation panels was to gain important insights into which filtering criteria were needed for generating high percentages of true SNP. Given the biased or skewed nature of SNPs selected for this purpose, therefore, we did not carry out a comprehensive population genetics analysis using the Sequenom genotypes. Future studies should utilize our optimized parameters to select representative SNP multiplex panels from the GBS data set. However, we did observe that the Sequenom genotypes showed a close relationship between D&B and Harvest Select fish, confirming the putative origin of Harvest Select blue catfish from D&B stock. Additionally, genotyping indicated an ancestral relationship between Kentucky Lake and Mississippi River blue catfish. Although Kentucky Lake is now impounded by Kentucky Dam, migration and mixture among blue catfish in the Tennessee, Ohio and Mississippi River drainages would have commonly occurred, as the species is known to migrate up to several hundred kilometres during certain seasons (Lagler 1966; Graham 1999).

Combining the data from both multiplexes, a total of 68 SNPs (four failed SNPs) and 191 individuals from seven populations (five GBS populations and two additional populations) were genotyped on the MassARRAY platform (Table S6, Supporting information). For the 64 successfully amplified SNPs, the average successful call rate was 92.97% (SD ± 5.01%), while 52 SNPs had a >90% call rate among all individuals with no more than two alleles per SNP. Of the attempted 12 224 individual genotypes, 959 individual genotypes were missed, resulting in a successful call rate of 92.15% (Table 3).

Conclusion

Blue catfish are a key member of large river ecological communities in North America, as well as a valued commercial and recreational species. We describe here the use of GBS for rapid generation of SNP resources and population-level genotypes for blue catfish. This study represents, to our knowledge, the first application of GBS in a fish species, and the first large-scale genetic analysis of blue catfish. The GBS approach generated 4275 high-quality SNPs across individuals from five tested populations. Further refinement of screening parameters (e.g. read coverage, minor allele frequency, observed heterozygosity, individual coverage) to select GBS SNPs for multiplex analysis can ensure a high rate of validation and multiplex utility. In the future, additional SNPs may be harvested from the present data set as the blue catfish genome assembly improves or through the use of different screening criteria or approaches. Given the somewhat cloudy history of blue catfish management in the USA, definitive population genetics analysis of wild and domesticated blue populations to address questions of ecology, conservation and aquaculture in the future will require sampling and genotyping larger numbers of individuals.

Acknowledgements

Authors wish to thank Sharon Mitchell and Katie Hyma at Cornell's Institute for Genomic Diversity for their work on GBS sequencing and initial data analysis. We are grateful to public and private collaborators for their assistance in sample collection. The study was conducted with funding from the Alabama Agricultural Experiment Station as well as USDA-ARS CRIS 6402-31000-009-00D. The USDA is an equal opportunity provider and employer. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

References

Arterburn JE, Kirby DJ, Berry CR Jr (2002) A survey of angler attitudes and biologist opinions regarding trophy catfish and their management. *Fisheries*, **27**, 10–21.

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

Bartram BL (2010) Factors affecting blue catfish populations in Texas reservoirs. *American Fisheries Society Symposium*, **77**, 187–197.

Bosworth BG, Waldbieser GC (2014) General and specific combining ability of male blue catfish *Ictalurus furcatus* and female channel catfish *I. punctatus* for growth and carcass yield of their F1 hybrid progeny. *Aquaculture*, **420–421**, 147–153.

Campbell NR, Narum SR (2011) Development of 54 novel single-nucleotide polymorphism (SNP) assays for sockeye and coho salmon and assessment of available SNPs to differentiate stocks within the Columbia River. *Molecular Ecology Resources*, **11**, 20–30.

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA (2013) Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genetics & Genomes*, **9**, 1537–1544.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One*, **8**, e62137.

Dunham R, Smitherman R (1984) *Ancestry and Breeding of Catfish in the United States, Cir. 273*. Alabama Agricultural Experiment Station, Auburn, AL.

Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Eggleton MA, Schramm HL Jr (2004) Feeding ecology and energetic relationships with habitat of blue catfish, *Ictalurus furcatus*, and flathead catfish, Pylodictis olivaris, in the lower Mississippi River, USA. *Environmental Biology of Fishes*, **70**, 107–121.

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

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

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

Freamo H, O'Reilly P, Berg PR, Lien S, Boulding EG (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, **11**, 254–267.

Giudice JJ (1966) Growth of a blue X channel catfish hybrid as compared to its parent species. *The Progressive Fish-Culturist*, **28**, 142–145.

Glaubitz JC, Casstevens TM, Lu F *et al.* (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*, **9**, e90346.

Glodek G (1980) *Ictalurus furcatus* (*Lesueur*), blue catfish. In: *Atlas of North American Freshwater Fishes*. (eds Lee DS and five coeditors) pp. 439. North Carolina State Museum of Natural History, Raleigh.

Gompert Z, Forister ML, Fordyce JA *et al.* (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Molecular Ecology*, **19**, 2455–2473.

Graham K (1999) A review of the biology and management of blue catfish. *American Fisheries Society Symposium*, **24**, 37–49.

Grist JD (2002) *Analysis of A Blue Catfish Population in A Southeastern Reservoir: Lake Norman*. North Carolina State University, Raleigh, North Carolina.

Hallerman EM, Dunham R, Smitherman R (1986) Selection or drift–isozyme allele frequency changes among channel catfish selected for rapid growth. *Transactions of the American Fisheries Society*, **115**, 60–68.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, **10**, 639–650.

Kopecký D, Studer B (2013) Emerging technologies advancing forage and turf grass genomics. *Biotechnology advances*, **32**, 190–199.

Krück NC, Innes DI, Ovenden JR (2013) New SNPs for population genetic analysis reveal possible cryptic speciation of eastern Australian sea mullet (*Mugil cephalus*). *Molecular Ecology Resources*, **13**, 715–725.

Lagler KF (1966) *Freshwater Fishery Biology*. pp. 421. W. M. C. Brown Company, Dubuque, Iowa.

Langella O (1999) Populations 1.2. 30. Accessed at: http://bioinformatics.org/~tryphon/populations (last accessed May 2009).

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

Liu Z, Nichols A, Li P, Dunham R (1998) Inheritance and usefulness of AFLP markers in channel catfish (*Ictalurus punctatus*), blue catfish (*I. furcatus*), and their F1, F2, and backcross hybrids. *Molecular and General Genetics MGG*, **258**, 260–268.

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

Mickett K, Morton C, Feng J *et al.* (2003) Assessing genetic diversity of domestic populations of channel catfish (*Ictalurus punctatus*) in Alabama using AFLP markers. *Aquaculture*, **228**, 91–105.

Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Parra-Bracamonte GM, Sifuentes-Rincón AM, Rosa-Reyna XFDl, Arellano-Vera W, Sosa-Reyes B (2011) Inbreeding evidence in a traditional channel catfish (*Ictalurus punctatus*) hatchery in Mexico. *Electronic Journal of Biotechnology*, **14**, 11–11.

Perales-Flores LE, Sifuentes-Rincón AM, León FJ (2007) Microsatellite variability analysis in farmed catfish (*Ictalurus punctatus*) from Tamaulipas, Mexico. *Genetics and Molecular Biology*, **30**, 570–574.

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

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

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

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

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Research*, **10**, 516–522.

Schloesser RW, Fabrizio MC, Latour RJ *et al.* (2011) Ecological role of blue catfish in Chesapeake Bay communities and implications for management. *American Fisheries Society Symposium*, **77**, 369–382.

Simmons M, Mickett K, Kucuktas H *et al.* (2006) Comparison of domestic and wild channel catfish (*Ictalurus punctatus*) populations provides no evidence for genetic impact. *Aquaculture*, **252**, 133–146.

Slate J, Gratten J, Beraldi D *et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97–107.

Tan G, Karsi A, Li P *et al.* (1999) Polymorphic microsatellite markers in *Ictalurus punctatus* and related catfish species. *Molecular Ecology*, **8**, 1758–1760.

Van Bers N, Crooijmans R, Groenen M, Dibbits B, Komen J (2012) SNP marker detection and genotyping in tilapia. *Molecular Ecology Resources*, **12**, 932–941.

Van Tassell CP, Smith TP, Matukumalli LK *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Vandepitte K, Honnay O, Mergeay J *et al.* (2013) SNP discovery using Paired-End RAD-tag sequencing on pooled genomic DNA of *Sisymbrium austriacum* (Brassicaceae). *Molecular Ecology Resources*, **13**, 269–275.

Waldbieser G, Bosworth B (1997) Cloning and characterization of microsatellite loci in channel catfish, *Ictalurus punctatus*. *Animal Genetics*, **28**, 295–298.

E.P. and G.W. planned the project and provided funding. E.P. and C.L. wrote the manuscript. C.L. and G.W. carried out data analysis. G.W., B.B. and B.B. assisted in sample collection and preparation and provided critical editing of the manuscript. C.L., E.P. and W.T. carried out the SNP validation.

Data Accessibility

In addition to filtered SNP genotype data and multiplex primer and genotype data available in Tables S1–4 (Supporting information), raw Illumina reads have been archived at the NCBI SRA: SRX483820. A VCF file representing the raw SNP data is available under DRYAD accession: doi:10.5061/dryad.4nf1c.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Plot of STRUCTURE estimation of best *K*. Plot of deltaK (a statistic based on the rate of change in the log probability of data between successive *K* values and used to predict the real number of clusters) where the modal value of the distribution is

considered as the highest level of structuring, in our case three clusters.

Table S1 Details of the 4275 SNPs identified from GBS of Blue Catfish.

Table S2 Summary of biallelic SNPs identified from genotyping by sequencing (GBS) in blue catfish with criteria minimum minor allele frequency = 0.01, minimum minor allele count = 10 and coverage of ≥70% of individuals within each population.

Table S3 Gene-associated SNPs with F_{ST} values identified from genotyping by sequencing (GBS) in blue catfish.

Table S4 Details of the 2214 SNPs used for population genetic analysis with only a single SNP per genomic region (contig).

Table S5 Population, allele, read count, and primer sequence information for the two Sequenom multiplex panels.

Table S6 Sequenom-generated SNP calls for the two multiplexes run on 191 blue catfish individuals.