Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis

C. He*,‡, L. Chen*,†,‡, M. Simmons*, P. Li*, S. Kim* and Z. J. Liu*

*Department of Fisheries and Allied Aquacultures and Program of Cell and Molecular Biosciences, The Fish Molecular Genetics and Biotechnology Laboratory, Aquatic Genomics Unit, Auburn University, Auburn, AL 3684, USA. †Department of Biology, East China Normal University, Shanghai 200062, China.
‡The first two authors contributed equally

Summary

In this study, we identified putative SNP markers within genes by comparative analysis of expressed sequence tags (ESTs). Comparison of 849 ESTs from blue catfish (*Ictalurus furcatus*) with >11 000 ESTs from channel catfish (*I. punctatus*) deposited in GenBank resulted in the identification of 1020 putative SNPs within 161 genes, of which 145 were nuclear genes of known function. The observed frequency of SNPs within ESTs of the two closely related catfish species was 1.32 SNP per 100 bp. The majority of identified SNPs differed between the two species and, therefore, these SNPs are useful for mapping genes in channel catfish × blue catfish interspecific resource families. The SNPs that differed within species were also observed; these can be applied to genome scans in channel catfish resource families.

Keywords aquaculture, expressed sequence tag, fish, marker, single nucleotide polymorphism.

Introduction

Comparative genome analysis requires the mapping of a common set of molecular markers in different species. Although linkage maps containing microsatellite markers are useful for comparing closely related species, gene-derived type I markers are essential for studies of genome evolution. Large numbers of type II markers have been developed from catfish for gene mapping and analysis of quantitative trait loci (for review see Liu 2003), but only a limited number of polymorphic type I markers in catfish are available (Liu *et al.* 1999, 2001). The objective of this project was to determine whether SNP within coding genes can be efficiently identified by comparative analysis of expressed sequence tags (ESTs) from channel catfish (*Ictalurus punctatus*) and blue catfish (*I. furcatus*) from which interspecific resource families were made (Liu *et al.* in press).

The advantage of SNPs for linkage analysis is provided by their abundant availability and even distribution (Weiss 1998), thus ensuring construction of high-density maps (Lai *et al.* 1998). The SNPs are the most common type of genetic variation in humans, with about one SNP per 1300 bp (Chakravarti 1999). They are usually biallelic and thus are less informative than multi-allelic microsatellite markers, but this can be compensated for by their abundance (Wang *et al.* 1998) and the use of SNP haplotypes. Additionally, platforms for SNP genotyping provide a high level of automation (Stickney *et al.* 2002).

In humans, SNP markers can be screened in cyberspace (Picoult-Newberg *et al.* 1999; Cox *et al.* 2001) or using DNA chips (Wang *et al.* 1998), but these methods are not yet available in aquaculture species such as catfish. Recently, ESTs have been used to identify gene-specific SNPs in equine (Shubitowski *et al.* 2001) and swine (Fahrenkrug *et al.* 2002). Their work indicated that ESTs are a rich source for SNPs. Here, we report that comparative EST analysis is an efficient approach for the identification of SNPs, using an interspecific hybrid system of channel catfish and blue catfish from which backcross resource families were made (Liu *et al.* 1998, in press).

Address for correspondence

Z. J. Liu, Department of Fisheries and Allied Aquacultures, Auburn University, 203 Swingle Hall, Auburn, AL 36849, USA.
E-mail: zliu@acesag.auburn.edu

Accepted for publication 24 August 2003

Materials and methods

Tissue preparation and RNA isolation

One-year-old blue catfish (*I. furcatus*) were raised in troughs located in the hatchery of the Auburn University Fish Genetics Facility for 4 weeks prior to the initiation of the experiments. At the start of the experiment, tricaine methanesulphonate (MS222; Argent Chemical Laboratories,

Redmond, WA, USA) at 300 ppm was used to kill the fish. Head kidney (also known as the anterior kidney) tissues were collected and cut into small pieces. Pooled head kidney tissues from 15 fishes were rapidly frozen in liquid nitrogen, ground with a mortar/pestle, and then homogenized with a hand-held tissue homogenizer in RNA extraction buffer using the guanidium thiocyanate method (Chomczynski & Sacchi 1987). Poly(A)$^+$ RNA was purified from total cellular RNA using the Poly(A)$^+$ Pure kit (Ambion, Austin, TX, USA) according to the manufacturer's instructions.

Construction of the blue catfish head kidney cDNA library

A directional cDNA library of blue catfish head kidney was constructed using the pSPORT-1 SuperScript Plasmid Cloning System (Invitrogen, Carlsbad, CA, USA) following the manufacturer's instructions except that the library was electroporated into ElectroMax DH12S cells (Invitrogen). Over 7.5 million primary cDNA clones were obtained with an average insert size of 1.0 kb. The primary cDNA library was amplified once before colonies were picked for sequencing.

Plasmid preparation and sequencing

The plasmid cDNA library was plated to a density appropriate for picking individual colonies. Random clones were grown overnight in 1.5-ml LB medium in 12 × 75-mm culture tubes. Plasmid DNA was prepared by the alkaline lysis method (Sambrook *et al.* 1989) using Qiagen Spin Column Mini-plasmid kits (Qiagen Inc., Valencia, CA) (Kocabas *et al.* 2002). Three microlitres of plasmid DNA (about 0.5–1.0 μg) was used in sequencing reactions. Chain termination sequencing was performed using cycleSeq-farOUT^{TM} polymerase (Display Systems Biotech, Vista, CA, USA). The polymerase chain reaction (PCR) profile was: 95 °C for 30 s, 55 °C for 40 s, 72 °C for 45 s for 30 cycles. An initial 2-min denaturation at 96 °C and a 5-min extension at 72 °C were used. Sequences were analysed on an automatic LI-COR DNA Sequencer Long ReadIR 4200 (LI-COR, Lincoln, NE) or LI-COR DNA Analyzer Gene ReadIR 4200.

Bioinformatic analysis

The BLAST searches were conducted to determine gene identities. Procedures for establishing orthologs were the same as previously described (Cao *et al.* 2001; Karsi *et al.* 2002). After the BLAST searches, putative gene identities were determined. All ESTs that were not identified as orthologs of known genes were designated as unknown EST clones. The EST sequences were then submitted to the dbEST database. The EST cluster analysis was conducted to identify the contigs using ContigExpress program of the Vector NTI software package. A list of blue catfish consensus transcripts was generated and used to search the dbEST to identify SNPs. The program BLASTN was used to search dbEST against 'EST_others'. All parameters were set at default except that the descriptions were limited to 50 and the alignments were limited to 10. The 'query-anchored with identities' function was selected to produce multiple sequence alignments anchored with the sequence coordinations of the query sequence. The search results were saved and SNPs were identified by visual inspection of the sequence alignments. The position of each SNP was given as a position on the reference sequence.

PCR amplification of the genomic segments containing SNPs

The PCR primers were designed using the OLIGO software package provided by Pyrosequencing, Inc. (Pyrosequencing AB, Uppsala, Sweden). Exon/intron border sequences (AG/GT) were inspected such that no primer spanned two exons. The PCRs of 50 μl were carried out in 50 mM KCl, 10 mM Tris (pH 9.0 at 25 °C), 0.1% Triton X-100, 0.25 mM each of dNTPs, 1.5 mM $MgCl_2$, 20 μM of the primers, about 500 ng genomic DNA pooled from 20 individuals and 2.5 units of *Taq* DNA polymerase. For loci that were relatively difficult to amplify, the FastStart PCR kit was used (Roche Applied Science, Indianapolis, IN, USA). One of the PCR primer pair was labelled with biotin for further analysis of SNPs using pyrosequencing (Sigma-Genosys, The Woodlands, TX, USA). The temperature profile was 94 °C for 30 s, 55 °C for 1 min and 72 °C for 2 min, for 35 cycles. The PCR products were run on agarose gels to determine yield.

Pyrosequencing analysis

The SNPs were analysed on a Pyrosequencer (Pyrosequencing, Inc., Westborough, MA, USA). The PCR product was immobilized, and single-strand isolation was performed with Dynabead M-280 Strepavidin (Dynal Biotech, Oslo, Norway) as described (Pyrosequencing AB). Pyrosequencing was performed at 28 °C in a total volume of 50 μl in an automated 96-well Pyrosequencer using PSQ SNP 96 enzymes and substrate (PSQ 96 SNP reagent kit, Pyrosequencing, Uppsala, Sweden) with cyclic dispersion of the nucleotides. The resulting sequences were first analysed automatically by the SNP evaluation software, and then manually by visual inspection of each pyrogram.

Results and discussion

Blue catfish ESTs

In order to generate EST sequence data for blue catfish, 849 clones from a head kidney cDNA library were end-sequenced. All sequences were deposited in GenBank with accession numbers of BQ096608–BQ097456. Clustering

analysis indicated that the 849 blue catfish ESTs represented 626 consensus transcripts, of which 316 were known genes and 310 were unknown EST clones. Sequences were annotated based on BLASTX analysis.

Putative SNP markers

Of the 626 consensus blue catfish transcripts, 170 had channel catfish counterparts in dbEST and 161 of these contained at least one SNP in the analysed region of the reference sequence (see Supplementary material, Table S1). The remaining nine genes were identical between the two species across the analysed region. Of those containing SNPs, 146 were transcripts of known genes and 15 were unknown. One of the 146 known genes was from mitochondria, leaving a total of 160 nuclear genes harbouring SNPs that differed between channel and blue catfish. The genes and locations of the interspecific SNPs are summarized in Supplementary material (Table S1).

Comparative EST analysis is an efficient approach for the identification of putative type I SNPs. Clearly, interspecific SNPs can be identified by targeted sequencing of PCR amplicon, i.e. amplification of the counterparts of selected genes followed by sequencing (Fahrenkrug *et al.* 2002). However, we had two objectives in this project: to identify interspecific SNPs and to expand catfish EST resources. The random EST sequencing approach used here should be very effective for SNP identification among highly expressed genes because highly abundantly expressed genes are represented as a high proportion of the cDNA library.

The observed interspecific SNP rate and types of base substitutions

Channel catfish and blue catfish are closely related species within the genus *Ictalurus*. An analysis of 161 genes indicated that overall the SNP rate between the genes of the two species was 1.32 SNP per 100 bp (Fig. 1). Channel catfish and blue catfish share an average 98.7% identity across the analysed genes. The majority of genes had an SNP rate between 0.5 and 2.0 per 100 bp, but this value ranged from 0 to 7.9 SNPs per 100 bp.

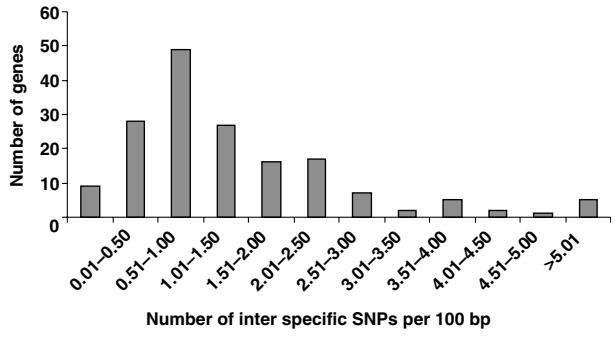

Figure 1 Distribution of SNPs (per 100 bp) within genes.

Across the 1020 SNPs within the 160 nuclear genes harbouring interspecific SNPs, the number of transitions (50.9%) was comparable with that of transversions (49.1%). However, the base substitution rate in transitions was more than twice the rate in transversions. This is because the transitions involved only two types of base substitutions: A/G (21.1%) and C/T (29.8%), whereas transversions involved four types of base substitutions: A/C (15.7%), A/T (13.2%), G/T (9.1%), and G/C (11.1%).

Applicability of the SNPs as markers for genome mapping in catfish

As previously reported, we have made reciprocal resource families by crossing F_1 (channel catfish female × blue catfish male) with either channel catfish or blue catfish (Liu *et al.* 1998, in press). To evaluate heterozygosity at SNP sites, we used a pooling/pyrosequencing strategy (Gruber *et al.* 2002) to determine the status of polymorphism in the resource families. Pooled DNA samples of 16–20 fishes were used for PCR amplification of SNP-containing fragments. In the backcross progeny, only two genotypes were expected: homozygous to the allele of the backcross parent or heterozygous. Among the progeny of two channel catfish backcross families (F_1–2 × channel catfish-1 and F_1–2 × channel catfish-6), the ratio of heterozygous individuals to homozygous individuals should be 1 : 1. Therefore, one base composition at the SNP site should be 75% for the backcross parent allele, and 25% for the alternative allele. This is what we observed with pyrosequencing, demonstrating that pyrosequencing with pooled samples was sufficient to assess SNPs for the purpose of genome mapping.

Of 20 randomly selected SNP sites, 11 SNPs segregated within one resource family (F_1–2 × channel catfish-6), while the remaining nine SNPs were homozygous within this resource family (see Supplementary material, Table S2). However, it should be noted that sequence variation was observed in many SNP sites within channel catfish as well. These SNPs should be useful in channel catfish resource families, but the polymorphisms need to be determined within each resource family. It is currently unknown which SNPs were only interspecific.

In summary, this research demonstrated that comparative EST analysis is an efficient approach for the development of SNPs in interspecific hybrid systems. Many SNPs identified from this work were species markers and therefore should be useful for mapping in interspecific hybrid resource families.

Acknowledgements

This project was supported by a grant from USDA NRI Animal Genome Basic Genome Reagents and Tools Program (USDA/NRICGP 2003-35205-12827). We appreciate the support of Auburn University, Department of Fisheries

and Allied Aquacultures, College of Agriculture, and the Vice President for Research for their matching funds to USDA NRI Equipment Grants (98-35208-6540, 99-35208-8512).

Supplementary material

The following material is available from: http://www.blackwellpublishing.com/products/journals/suppmat/AGE/AGE1054/AGE1054sm.htm

Table S1: Putative SNPs identified by comparing EST sequences of blue catfish (*I. furcatus*) with those of channel catfish (*I. punctatus*).

Table S2: PCR and pyrosequencing primers used to determine informativeness of 20 SNPs within a resource family F_1–2 × channel catfish-6.

References

Cao D., Kocabas A., Ju Z., Karsi A., Li P., Patterson A. & Liu Z. J. (2001) Transcriptome of channel catfish (*Ictalurus punctatus*): initial analysis of genes and expression profiles from the head kidney. *Animal Genetics* **32,** 169–88.

Chakravarti A. (1999) Population genetics-making sense out of sequence. *Nature Genetics* **21,** 56–60.

Chomczynski P. & Sacchi N., (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry* **162,** 156–9.

Cox D., Boillot C. & Canzian F. (2001) Data mining: efficiency of using sequence databases for polymorphism discovery. *Human Mutation* **17,** 141–50.

Fahrenkrug S.C., Freking B.A., Smith T.P., Rohrer G.A. & Keele J.W. (2002) Single nucleotide polymorphism (SNP) discovery in porcine expressed genes. *Animal Genetics* **33,** 186–95.

Gruber J.D., Colligan P.B. & Wolford J.K. (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using pyrosequencing. *Human Genetics* **110,** 395–401.

Karsi A., Cao D., Li P., Patterson A., Kocabas A., Feng J., Ju Z., Mickett K. & Liu Z.J. (2002) Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. *Gene* **285,** 157–68.

Kocabas A., Li P., Cao D., Karsi A., He C., Patterson A., Ju Z., Dunham R. & Liu Z.J. (2002) Expression profile of the channel catfish spleen: analysis of genes involved in immune functions. *Marine Biotechnology* **4,** 526–36.

Lai E., Riley J., Purvis I. & Roses A. (1998) A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54,** 31–8.

Liu Z.J. (2003) A review of catfish genomics: progress and perspectives. *Comparative and Functional Genomics* **4,** 259–65.

Liu Z.J., 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* **258,** 260–8.

Liu Z.J., Karsi A. & Dunham R.A. (1999) Development of polymorphic EST markers suitable for genetic linkage mapping of catfish. *Marine Biotechnology* **1,** 437–47.

Liu Z.J., Li P., Kocabas A., Ju Z., Karsi A., Cao D. & Patterson A. (2001) Microsatellite-containing genes from the channel catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B. *Biochemical and Biophysical Research Communications* **289,** 317–24.

Liu Z.J., Karsi A., Li P., Cao D. & Dunham R. An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family. *Genetics* (in press).

Picoult-Newberg L., Ideker T.E., Pohl M.G., Taylor S.L., Donaldson M.A., Nickerson D.A. & Boyce-Jacino M. (1999) Mining SNPs from EST databases. *Genome Research* **9,** 167–74.

Sambrook J., Frisch E.F. & Maniatis T. (1989) *Molecular Cloning.* A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Shubitowski D.M., Venta P.J., Douglass C.L., Zhou R.X. & Ewart S.L. (2001) Polymorphism identification within 50 equine gene-specific sequence tagged sites. *Animal Genetics* **32,** 78–88.

Stickney H.L., Schmutz J., Woods I.G., Holtzer C.C., Dickson M.C., Kelly P.D., Myers R.M. & Talbot W.S. (2002) Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. *Genome Research* **12,** 1929–34.

Wang D.G., Fan J.B., Siao C.J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **15,** 1077–82.

Weiss K.M. (1998) In search of human variation. *Genome Research* **8,** 691–7.