Mar Biotechnol (2012) 14:672–680
DOI 10.1007/s10126-012-9449-0

SHORT COMMUNICATION

Use of Comparative Genomics to Develop EST-SSRs for Red Drum (*Sciaenops ocellatus*)

Christopher M. Hollenbeck · David S. Portnoy · John R. Gold

Received: 23 January 2012 / Accepted: 3 March 2012 / Published online: 19 April 2012

Abstract Microsatellites physically linked to expressed sequence tags (EST-SSRs) are an important resource for linkage mapping and comparative genomics, and data mining in publicly available EST databases is a common strategy for EST-SSR discovery. At present, many species lack species-specific EST sequence data needed for the efficient characterization of EST-SSRs. This paper describes the discovery and development of EST-SSRs for red drum (*Sciaenops ocellatus*), an estuarine-dependent sciaenid species of economic importance in the USA and elsewhere, using a phylogenetically informed, comparative genomics approach to primer design. The approach entailed comparing existing genomic resources from species closely allied phylogenetically to red drum, with resources from more distantly related outgroup species. By taking into account the degree to which flanking regions are conserved across taxa, the efficiency of PCR primer design was increased greatly. The amplification success rate for primers designed for red drum was 100 % when using EST libraries from confamilial species and 92 % when using an EST library from a species in the same suborder. The primers developed also amplified EST-SSRs in a wide range of perciform fishes, suggesting potential use in comparative genomics. This study demonstrates that EST-SSRs can be efficiently developed for an organism when limited species-specific data are available by exploiting genomic resources from well-studied species, even those at extended phylogenetic distances.

C. M. Hollenbeck (✉) · D. S. Portnoy · J. R. Gold
Department of Wildlife and Fisheries Sciences,
Center for Biosystematics and Biodiversity,
Texas A&M University,
TAMU 2258,
College Station, TX 77843-2258, USA
e-mail: chollenbeck07@tamu.edu

Keywords EST-SSRs · Data mining · Red drum · Comparative genomics

Introduction

The availability of DNA sequence data in public databases has increased dramatically over the last decade. Consequently, "data mining" from these databases is being used to address a variety of biological questions (Nagaraj et al. 2007). Expressed sequence tags (ESTs) in particular have become a useful resource for data mining as a result of their abundance and availability in a diversity of biota (Ellis and Burke 2007). Because microsatellites (SSRs) are present in a significant percentage of ESTs, the discovery and design of EST-SSRs by data mining EST libraries has become a common strategy (Bouck and Vision 2007). EST-SSRs are desirable markers because they are abundant and evenly dispersed across genomes (Kantety et al. 2002; Ju et al. 2005), and use of publicly available EST libraries can bypass the time and expense of enriched library design (Squirrell et al. 2003). In addition, because ESTs represent cDNA copies of expressed sequences (Adams et al. 1991), EST-SSRs are tightly linked to functional coding genes whose identity often can be ascertained with a BLAST search. Finally, because primers are designed in relatively conserved coding regions, EST-SSRs display higher cross-species transferability than non-EST-derived SSRs (so-called genomic microsatellites; Coulibaly et al. 2005; Varshney et al. 2005).

At present, however, most organisms lack the species-specific EST sequence data necessary for characterizing EST-SSRs. This problem can be circumvented by screening EST libraries of closely related species for SSRs and designing a large number of primers to cross-amplify loci in the species of interest (Yue et al. 2004; Li and Li 2008; Zhou et al. 2008). However, such "shotgun" approaches to EST-SSR cross-amplification commonly result in <50 % of markers successfully amplifying in the species of interest. Among the most important factors affecting cross-amplification success is the presence of primer–site mismatches between the species from which the ESTs are derived and the target species (Housley et al. 2006). The design of primers with binding sites that are highly conserved across taxa can therefore maximize cross-species transferability and amplification success (Dawson et al. 2010). The wealth of EST and genomic sequence data available in public databases makes it possible to employ a comparative genomics approach to primer design which considers the phylogenetic relationships between the species involved, thereby maximizing the efficiency of cross-amplifying EST-SSRs in a target species. The logic behind such a methodology is as follows: if primer binding sites are conserved between a species in a given clade (the ingroup) and a species in a different clade (the outgroup), those sites are likely to be conserved among other species within the ingroup. Specifically, by aligning EST-SSR sequences from an organism in the ingroup to the genomic sequences of an organism in the outgroup and designing primers only in conserved portions of the flanking regions, it should be possible to achieve a high degree of amplification success in another species in the ingroup (i.e., the target species).

In this paper, we exploit this methodology by designing EST-SSRs for a target species, red drum (*Sciaenops ocellatus*), using the EST libraries of phylogenetically related species. Red drum is an estuarine-dependent, economically important sciaenid species that is aquacultured globally and is part of an important recreational fishery in the USA, and for which there are currently no publicly available EST sequences. In addition, we report on a software tool designed for the automation of this process.

Materials and Methods

All EST sequences were downloaded from GenBank dbEST (http://www.ncbi.nlm.nih.gov/nucest). Downloaded EST sequences were placed into two ingroups in order to assess the phylogenetic distance at which the method could be employed efficiently. The first, more exclusive ingroup consisted of EST sequences from two species, yellow croaker (*Larimichthys crocea*) and miiuy croaker (*Miichthys miiuy*), in the family Sciaenidae, to which red drum also belongs. The second ingroup consisted of EST sequences from a moronid species, European sea bass (*Dicentrarchus labrax*), which is in the suborder Percoidei, as is the red drum. The outgroup chosen for both levels of comparison was the Nile tilapia, *Oreochromis niloticus*. This perciform species was selected as it represented the closest outgroup species (suborder Labroidei) with a large amount of available genome sequence data.

EST sequences were screened for contaminating vector and linker sequences using NCBI's VECSCREEN (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) and trimmed to remove regions with significant hits to known vectors. Sequences were then trimmed to remove poly-T and poly-A regions at the 5′ and 3′ ends, respectively. Sequences that were <100 bp in length after trimming were removed from further analysis. To eliminate redundancy, the remaining sequences were assembled for each species into a set of consensus sequences using the program CAP3 (Huang and Madan 1999). A custom Perl script was used to screen the consensus sequences for microsatellite motifs using the minimum repeat criteria of 5, 5, and 4 for di-, tri-, and tetranucleotide repeats, respectively. Sequences containing microsatellite motifs were then aligned, using the BLASTN program within NCBI's BLAST+ suite (Camacho et al. 2009), to whole-genome shotgun (WGS) sequences of the Nile tilapia. Alignments with an *e* value below a threshold of 10 and containing a region of repetitive sequence were selected as candidates for primer design. Primers were designed manually using the program PRIMER3 (Rozen and Skaletsky 2000; http://frodo.wi.mit.edu/). Only highly conserved flanking regions in which no indels and a maximum of two nucleotide substitutions were present were selected as candidate regions for primer placement. Twenty-five primer pairs were designed using sequences from both ingroups. Steps in the process from trimming ESTs to alignment with the *O. niloticus* genome were automated using a custom tool (available upon request from CMH), written in Perl, that utilizes a number of modules from the BioPerl project (Stajich et al. 2002).

Polymerase chain reaction (PCR) amplification success and microsatellite polymorphism in red drum were assessed initially using four red drum obtained from the Gulf of Mexico near Corpus Christi, Texas. An additional 16 red drum from the same location were genotyped to assess allele frequencies at all polymorphic microsatellites. Cross-species amplification was assessed using four individuals each of six additional perciform species: five from the suborder Percoidei—spotted sea trout (*Cynoscion nebulosus*) in the family Sciaenidae, red snapper (*Lutjanus campechanus*) in the family Lutjanidae, coney (*Cephalopholis fulva*) in the family Serranidae, greater amberjack (*Seriola dumerili*) in the family Carangidae, and cobia (*Rachycentron canadum*) in the family Rachycentridae—and one from the suborder Scombroidei—Serra Spanish mackerel (*Scomberomorus brasiliensis*). Genomic DNA was extracted from fin clips using a modified Chelex extraction protocol (Estoup et al. 1996). Amplifications were performed using the "tailed" protocol described in Karlsson et al. (2008). PCR products were electrophoresed and visualized on 6 % polyacrylamide gels using an ABI Prism 377 automated sequencer (Applied Biosystems). Allele calling was performed manually using GENOTYPER 2.5 (Perkin Elmer) and GENESCAN 3.1.2 (Applied Biosystems). The number of alleles and observed and expected heterozygosity were calculated, and exact tests of conformity to the expectations of Hardy–Weinberg equilibrium (HWE) were performed using GENEPOP 007 (Rousset 2008). The presence of scoring errors, large-allele dropout, and null alleles was assessed using MICRO-CHECKER (van Oosterhout et al. 2004). Polymorphic EST-SSRs were screened for putative function using a BLASTN and BLASTX search against the NCBI non-redundant (nr) database for nucleotides and proteins, respectively.

Results

After trimming ESTs, removing small sequences, and assembling sequences to eliminate redundancy, a total of 3,912 and 24,794 sequences remained from the two sciaenids (yellow croaker and miiuy croaker) and the European seabass, respectively. A total of 420 microsatellites (7.2 % of all non-redundant ESTs) were detected in sequences from the two sciaenids, and of these, 173 sequences had an acceptable BLAST hit against *O. niloticus* WGS sequences. All 25 primer pairs (100 %) amplified a scoreable product in red drum and 13 of 25 (52 %) were polymorphic (Table 1). For one microsatellite (*Soc_Mmi*07), there were several amplified products, some of which may have been artifacts; the remaining 11 were monomorphic. A total of 2,900 microsatellites (5.2 % of all non-redundant ESTs) were detected in sequences from European seabass. Of these, 1,445 had an acceptable BLAST hit against *O. niloticus* WGS sequences. Twenty-three of the 25 primers (92 %) amplified in red drum, and 16 (64 %) were found to be polymorphic (Table 1). Among all microsatellites, the number of alleles ranged from 2 (*Soc_Lcr*09, *Soc_Lcr*12, *Soc_Mmi*05, *Soc_Dla*18, *Soc_Dla*18, *Soc_Dla*22) to 17 (*Soc_Lcr*03), and gene diversity ranged from 0.05 (*Soc_Dla*18) to 0.934 (*Soc_Lcr*03). Significant departures from the expectations of HWE prior to sequential Bonferroni correction (Rice 1989) were found for 7 of 29 microsatellites; genotypes at only one of those microsatellites (*Soc_Dla*13) deviated significantly from the expectations of HWE following correction. Analysis with MICRO-CHECKER indicated the possible occurrence of null allele(s) at four microsatellites (*Soc_Lcr*14, *Soc_Dla*09, *Soc_Dla*13, and *Soc_Dla*21). A function was assigned to 17 of the 29 polymorphic EST-SSRs and to 11 of the 18 monomorphic EST-SSRs based on BLASTN or BLASTX searches (Table 1). Because these loci were not sequenced in the target species, the assigned functions reported are putative. Of the 28 microsatellites that could be assigned a putative homology, 17 were located in 3′ untranslated regions (UTRs), 3 were located in 5′ UTRs, and 4 were located in amino acid coding regions. All microsatellites located in coding regions were trinucleotide repeats. The remaining four microsatellites had significant BLAST hits to probable pseudogenes, genomic introns, or unidentified mRNAs. Amplification success in other species, defined as the presence of one or more bands scored on a polyacrylamide gel near the approximate size of the fragment observed in red drum, ranged from 76 % in *R. canadum* to 100 % in *C. fulva* for the 29 polymorphic microsatellites assayed (Table 2). The percentage of polymorphic EST-SSRs for each species ranged from 21 % in *S. dumerili* to 66 % in *C. nebulosus*. One microsatellite, *Soc_Lcr*07, was polymorphic in all species investigated.

Discussion

The results of this study demonstrate that a comparative genomics approach to primer design can be used to efficiently screen for and develop EST-SSRs for a species of interest from EST libraries of related species. Amplification success was shown to decrease only slightly when using ESTs from a species in a more distant taxon (different family, same suborder) compared to ESTs mined from confamilial species. The total amplification success rate observed in this study (96 %) is highly efficient, even when compared to EST-SSR mining studies that have employed species-specific ESTs (Wang and Guo 2007; Yu and Li 2008; Vogiatzi et al. 2011). While the proportion of polymorphic markers ultimately obtained is difficult to control, the approach applied in this study increases the ratio of polymorphic markers to the primers tested by minimizing the number of primers that fail to amplify a product altogether. In total, 29 of 50 designed primer pairs amplified a scoreable, polymorphic microsatellite in red drum. This was accomplished using <50 % of the available sciaenid EST-SSRs with an acceptable BLAST alignment against the *O. niloticus* genome and <5 % of the *D. labrax* EST-SSRs with an acceptable alignment. This suggests that the same technique could be used to design many additional markers using the same EST libraries.

A principal advantage of this approach is that it can theoretically be extended to incorporate ESTs at arbitrarily large taxonomic distances by making the ingroup more inclusive and selecting an appropriate outgroup. The ESTs utilized in this study represent only ~30 % of all percoid ESTs and only ~15 % of perciform ESTs that could be subject to the same search strategy. The choice of the outgroup genome is another factor that can increase the ultimate yield. While *O. niloticus* was an appropriate and

Table 1 Summary information for 47 EST-SSRs designed for *S. ocellatus* from *L. crocea*, *M. miiuy*, and *D. labrax* ESTs

Accession	Marker	Motif	Primers	Size range (bp)	#A	H_E	P_{HW}	F_{IS}	Putative homology	*E* value
CX348392.1	*Soc_Lcr*01	$(AC)_7$ G$(CA)_8$TG$(CA)_5$	F: GCCTGCTTGAATTGTGTTGC R: CATATAATCTGTGGGCAGGAAG	248–268[b]	6	0.605	0.024[c]	0.130	GATA-binding protein 3[e]	4.00E−20
CX348925.1	*Soc_Lcr*03	$(TC)_5$ C$(CT)_6$A$(TC)_6$	F: TCTTTTCATTTGGTTCATACAAGC R: CCAGCCTGTTTAACAAGTAGGTC	161–211	17	0.934	0.033[c]	0.090		
CX348360.1	*Soc_Lcr*04	$(CT)_8$ $N_7(CT)_5$	F: CTGGACTCCAAAGTTGAGTGG R: GTGACAGACAATTGTGCTCTTTT	210–220	3	0.271	1.000	−0.107	Guanine nucleotide binding protein beta polypeptide[f]	3.00E−90
CX348626.1	*Soc_Lcr*07	$(AC)_{16}$	F: CATGGTCATTAGCAAGTAGAGTTCA R: AGAAGCACCGATTTGGTCTG	273–277[b]	5	0.459	0.060	0.238	Hypothetical protein LOC100342906[g]	2.00E−18
CX348990.1	*Soc_Lcr*08	$(TG)_5$ C$(GT)_{13}$	F: TGCTGGTCTCGACCTTAATTG R: GACATTTATTTAAAAACATTTGTTCAG	357–375	9	0.884	0.047[c]	0.039	Type III iodothyronine deiodinase[h]	8.00E−91
EV413954.1	*Soc_Lcr*09	$(TG)_2$ C$(TG)_2$TA$(TG)_4$TC $(TG)_5$	F: TTGAGCTGTATTTCATCAAAGC R: CCTTACTTCATGCACACATGC	241–242[b]	2	0.511	0.394	0.217	Apolipoprotein A-I[i]	5.00E−97
EB643369.1	*Soc_Lcr*12	$(CAT)_9$	F: TCTGGAAATGTTGGATAAATG R: GGCATTAGTGACTTGCGATTG	167–173	2	0.405	0.256	−0.357		
EB643365.1	*Soc_Lcr*14[a]	$(AC)_{27}$	F: TTGGCATAAAGTTAAACCATTCAG R: ACTCATCCTGTGACATGAACTC	201–239	9	0.858	0.106	0.242		
GW667871.1	*Soc_Mmi*04	$(AC)_6$ T$(CA)_7$	F: AGGCGCCATTAATTTGAGTG R: TCTGTGGGTATGTGCGTGTT	151–155	3	0.538	1.000	0.071		
GW670256.1	*Soc_Mmi*05	$(TTTC)_{11}$	F: GGCAGAAGTGGAACTGTTGTAG R: TCACTCCAGTCCAGTGTTTGA	318–322	2	0.263	0.354	0.240		
GW670585.1	*Soc_Mmi*06	$(TC)_7(AC)_8$	F: TGCCTTATGTAAGTGGCCTTG R: ACTGTATGGATCAGAGCCCTTT	156–162[b]	5	0.58	0.764	−0.034	Krueppel-like factor 6 putative[j]	8.00E−32
GW672095.1	*Soc_Mmi*10	$(GAT)_4$TG $(GAT)_5$GAC $(GAT)_3$	F: AAGTGTGCCTGCCTGTGC R: GCAAATAGTAGGTGTATTGCAACG	232–241	4	0.601	1.000	−0.081		
GW672243.1	*Soc_Mmi*11	$(AC)_9$	F: AGTGTCTGCTGGATCACTATGC R: TGCGAGGAGAGACATTTGG	180–194	4	0.629	0.273	0.126	Disabled homolog 1a[k]	6.00E−21
FM019990.1	*Soc_Dla*01	$(TGA)_{11}$	F: AGCTCACCTGGCGCTGAC R: TCCAAATGAGTCGTGTGCTTG	330–352[b]	15	0.925	0.432	0.027	Unidentified cDNA[l]	6.00E−05
FM004106.1	*Soc_Dla*02	$(CAG)_7$	F: GCCACAATCAGCAACTGAAC R: AATTCACGGATCTGGTCATTC	240–249	4	0.684	0.168	0.196	Uracil phosphoribosyltransferase[m]	2.00E−161
FP242838.1	*Soc_Dla*03	$(GT)_4$GAA $(TG)_5$	F: CTAGGCTGTAATAATGTTCACTGT R: GTATCCCTAAGGAACCAAGG	165–168[b]	3	0.229	1.000	−0.092	Arginine-glutamic acid dipeptide (RE) repeats[n]	1.00E−33
FP239107.1	*Soc_Dla*04	$(GT)_{11}$	F: GTGCCTTTTATATTTCTCTTGACAC R: CACATGGTAGCAAGTGTATTTTTG	256–269	6	0.755	0.477	−0.059		
FM001652.1	*Soc_Dla*09[a]	$(GA)_{23}$	F: AAGCACAGCTTTGTAAATGCAC R: TGACAGGACATGGCTTATTATCA	373–389	8	0.845	0.014[c]	0.295	CLIP-associating protein 2[j]	1.00E−18
FM019804.1	*Soc_Dla*10	$(TC)_6$	F: TTGCAATTTGAAGGATAAAGC R: CTGCACACATGAATCACTCC	308–332	11	0.850	0.206	0.060	Histone deacetylase 9[n]	5.00E−06
FM020143.1	*Soc_Dla*12	$(TTA)_6$	F: GAAATGATTATTGAACTACGCAAC R: GCCAATGGGATCCAAGC	116–122	3	0.145	1.000	−0.036		
FM001773.1	*Soc_Dla*13[a]	$(CA)_6$CG $(CA)_{12}$	F: GGTGCCGACAAGGTCAAC R: AAGAGAAGGTGAACTGAGAGTAAGG	379–387	4	0.391	0.0001[d]	0.749	ATPase asna1[o]	9.00E−80

Table 1 (continued)

Accession	Marker	Motif	Primers	Size range (bp)	#A	H_{E}	P_{HW}	F_{IS}	Putative homology	*E* value
FM028355.1	*Soc_Dla*14	$(GA)_{14}$	F: CGTTGGACCTGGAGATGG R: CACAGTGACTGGTTGAACTGC	269–275	3	0.099	1.000	−0.013	EGF-containing fibulin-like extracellular matrix protein[o]	2.00E−81
FM001243.1	*Soc_Dla*16	$(AC)_5$ G$(CA)_4$ T$(AC)_4$	F: TCTCAGCACATTTGTACAGTAGTTTG R: ACTGCATCTGCACAATAAGTGAC	281–297	8	0.753	0.070	−0.133		
FM000143.1	*Soc_Dla*17	$(AG)_{11}$	F: GACGCGAGCGAAAGGAAC R: CTCCAGCATCCAGTCCTCTG	222–226	3	0.529	0.353	0.249		
FM025924.1	*Soc_Dla*18	$(TC)_2$A$(CT)_9$	F: TTACTTCACTACTCTAAAGGAACAAC R: TCACATAGTAAAACGACGACAG	296–298	2	0.050	n/a	0.000	Parvalbumin[p]	5.00E−62
FM026271.1	*Soc_Dla*21[a]	$(TG)_2$ C$(GT)_6$ $N_{11}(GT)_4$ $(GAGT)_2$ $(GT)_3$	F: TGTCTGTATACTGTATTTTAGAGGTCA R: CAAACCTTGTTGCACCAGAA	336–340	3	0.478	0.008[c]	0.484		
FM009184.1	*Soc_Dla*22	$(CT)_6$	F: TGTAGCTCTGCAGCCTCCTC R: GTGATGAACGAGCCCTTAGTG	153–155	2	0.467	0.043[c]	−0.520		
FK940711.1	*Soc_Dla*23	$(GAG)_4$GAA $(GAG)_2$	F: GACGACGTCAAGGTGGAG R: TGCATTCTCTACAGTTACCAAATG	284–286	2	0.097	1.000	−0.027	Vacuolar protein sorting-associated protein 29[j]	0.00E+00
FM011451.1	*Soc_Dla*24	$(TC)_{12}$	F: CAGATGCAAATGTTTGTGTAGC R: AAAATTAGAACACCTTAAGCAGAGG	182–188	4	0.510	0.146	0.121	Myristoylated alanine-rich C-kinase substrate[j]	0.00E+00
Monomorphic EST-SSRs										
CX348550.1	*Soc_Lcr*02	$(CT)_7$TTC $(TTTC)_3$ $(TC)_7(T)_3$ $(CT)_2$	F: GCCAAGTTCTTCTCAGCACTC R: GGACTTGTAACTTGTGTCATTGC	241					Basic leucine zipper transcriptional factor ATF-like[o]	1.00E−69
CX348556.1	*Soc_Lcr*05	$(AC)_6$	F: CACGTTCTTCTCAGCACTCG R: CTTGTGTCATTGCTGTCAGAG	228					Basic leucine zipper transcriptional factor ATF-like[o]	1.00E−74
CX348612.1	*Soc_Lcr*06	$(TC)_3$TT $(TC)_8$	F: AGCACCACCTGGGATTACTGG R: CCCATTCTGGACAAGAGAAACC	373						
EV413959.1	*Soc_Lcr*10	$(AC)_5$AG $(AC)_4$AT $(AC)_2$AG $(CA)_2$	F: CTCCCAACCCTGTTCCTG R: TTGAGCTGTATTTCATCAAAGC	354					Clone apoa1_3 apolipoprotein A-I[q]	2.00E−155
EB643292.1	*Soc_Lcr*13	$(GT)_7$AT $(GT)_5$	F: CTCTTTAAGCCTGAGTCTCTGAGG R: ACAGCCGAGAAGCTTTAACG	394						
GW668767.1	*Soc_Mmi*01	$(GT)_6$	F: CCTCTCTTTCCAGCTAAGTATC R: GTGAATGAAATCAGTCTTTATCTG	239						
GW668773.1	*Soc_Mmi*02	$(CA)_7$	F: TAAGCCAAGTTCTTCTCAGC R: GACTTGTAACTTGTGTCATTGC	242					Basic leucine zipper transcriptional factor ATF-like[o]	2.00E−122
GW669238.1	*Soc_Mmi*03	$(TG)_{15}$	F: CTGACTCCTGACCATTTGTTATG R: CGATGACCACTGACCCAATAG	406					Gsh2, Pdgfra, and Kita, Kdrb, and Clock[r]	1.00E−88
GW670899.1	*Soc_Mmi*08	$(GAT)_6$	F: GGATGAGTGTGGAGCTACAGG R: CAGCTCATCACGGCTCAG	281					Nuclear factor erythroid 2-related factor 1-like[o]	2.00E−136
GW671772.1	*Soc_Mmi*09	$(CA)_6$AA $(CA)_4$	F: GGTTGCCTGAGCCGTAGG R: ACTAATGACTTGCGGCCTAGC	115						
GW672302.1	*Soc_Mmi*12	(CAG)CA (CAG)C $(CAG)_4$	F: AGCTGGGCATTGAAGAGC R: GTCGCACTGACTCGTCTCC	254					CCAAT/enhancer-binding protein beta 2[s]	0.00E+00
FK943283.1	*Soc_Dla*05	$(CAT)_2$CA $(CAT)_6$	F: CTGACTCGAGCGCATTTTTA R: TTGGAACGTGTTTTTATATTAGGG	239					Sodium/potassium-transporting ATPase subunit alpha-1-like[o]	5.00E−93
FP240767.1	*Soc_Dla*06	$(CA)_5$AA $(CA)_3$	F: ATCCCACAATGCTGTGAGC	170						

Table 1 (continued)

Accession	Marker	Motif	Primers	Size range (bp)	#A	H_E	P_{HW}	F_{IS}	Putative homology	*E* value
			R: GGATGGATGGAGACAATTAAGC							
FK943099.1	*Soc_Dla*07	$(GT)_4GC\ (GT)_3TT\ (GT)_3$	F: TGTGAAATCATTCCAGTAAACTATCTG R: TGACAAGTCCTGCAATTTAAAGC	195					Beta-synuclein putative[e]	0.00E+00
FK940130.1	*Soc_Dla*08	$(TG)_6$	F: AAATGAGCAAATCCTTCAGATTG R: CACATCACTTAAGGCTCATTTCTG	60						
FK940504.1	*Soc_Dla*15	$(GAA)_5$	F: CACATATGACTCTGCCTGCTG R: GGCTGTCTGTTCCTGCTG	346					C3orf54 putative[j]	4.00E−138
FM019892.1	*Soc_Dla*20	$(GGAT)_5$	F: ATCACGGTCGCCACATAATC R: GGCTGTCTGTTCCTGCTG	273						
FM019678.1	*Soc_Dla*25	$(AC)_6$	F: CACACCTGCGTATCTGACTGA R: CAACAGCCATAAATTTGAAAAGG	144					Ubiquitin-associated protein 2[f]	0.00E+00

EST-SSRs are named with the first letter of the genus and first two letters of the species name for the target species, followed by an underscore and the same pairing of letters for the species from which the EST sequence was taken. F_{IS} is Weir and Cockerham's (1984) *f*, as calculated in GENEPOP 007 (Rousset 2008)

[a] Possible null allele indicated by MICRO-CHECKER

[b] 1-bp shifts in allele size

[c] *P* value significant before correction

[d] *P* value significant after correction

[e] *O. mykiss*

[f] *D. labrax*

[g] *O. cuniculus*

[h] *H. trimaculatus*

[i] *O. fasciatus*

[j] *S. salar*

[k] *D. rerio*

[l] *T. nigroviridis*

[m] *A. fimbria*

[n] *G. gallus*

[o] *O. niloticus*

[p] *S. chuatsi*

[q] *M. saxatilis*

[r] *A. burtoni*

[s] *E. coioides*

Table 2 Cross-amplification of 29 EST-SSRs in six perciform fishes

Marker	*C. nebulosus*[a]	*C. fulva*[b]	*L. campechanus*[c]	*S. dumerili*[d]	*R. canadum*[e]	*S. brasiliensis*[f]
*Soc_Lcr*01	P	A	P	–	–	P
*Soc_Lcr*03	P	A	–	A	–	–
*Soc_Lcr*04	P	A	P	A	A	P
*Soc_Lcr*07	P	P	P	P	P	P
*Soc_Lcr*08	A	A	–	P	–	P
*Soc_Lcr*09	A	A	P	–	P	A
*Soc_Lcr*12	A	A	P	A	P	P
*Soc_Lcr*14	P	P	A	–	–	–
*Soc_Mmi*04	P	A	A	A	A	A
*Soc_Mmi*05	P	A	P	A	P	P
*Soc_Mmi*06	P	P	P	A	A	P
*Soc_Mmi*10	P	P	P	A	–	P
*Soc_Mmi*11	P	P	P	P	A	A
*Soc_Dla*01	A	A	A	A	A	A
*Soc_Dla*02	P	P	P	–	–	A
*Soc_Dla*03	A	A	P	P	P	P
*Soc_Dla*04	P	A	P	A	A	A
*Soc_Dla*09	A	P	P	A	A	A
*Soc_Dla*10	–	A	A	P	A	–
*Soc_Dla*12	P	P	P	A	A	P
*Soc_Dla*13	P	P	P	A	–	A
*Soc_Dla*14	A	A	A	A	A	P
*Soc_Dla*16	P	P	P	A	P	P
*Soc_Dla*17	A	A	P	A	A	A
*Soc_Dla*18	P	A	A	P	P	–
*Soc_Dla*21	A	P	A	A	A	P
*Soc_Dla*22	P	A	A	A	A	P
*Soc_Dla*23	P	P	A	A	A	P
*Soc_Dla*24	P	P	P	A	A	P

The presence of a "P" denotes scoreable, polymorphic amplification product; "A" denotes amplification success, but product is monomorphic or not easily scored; '–' denotes failed amplification

[a] Family Sciaenidae

[b] Family Serranidae

[c] Family Lutjanidae

[d] Family Carangidae

[e] Family Rachycentridae

[f] Family Scombridae

effective outgroup for this study, the WGS sequences currently available for this species represent an incomplete genome. The number of potential markers could therefore be increased using a more complete genome or, alternatively, multiple outgroup genomes. Although the results of this study indicate otherwise, it should be noted that ESTs utilized from species at greater phylogenetic distances from the target species may yield a lower percentage of conserved sequences (i.e., significant BLAST hits) and polymorphism, as both likely are a function of the time that a given region of DNA has been on a separate evolutionary trajectory (Primmer et al. 2005). The higher percentage of polymorphic markers designed from *D. labrax* ESTs is likely the result of having an abundance of available sequences and, thus, the option of using more selectivity in designing primers (favoring longer repeats, etc.).

The methodology applied in this study maximizes cross-amplification success by limiting primer–site mismatches between species. In addition, the method minimizes the number of failed amplifications resulting from primers placed on either side of an exon/intron splice site. Like nucleotide composition, conservation of splice sites between the groups can be inferred with the alignment of an EST to a contiguous genomic sequence of the outgroup. By strictly using EST-to-genome alignments to design primers, the likelihood of a splice site being located in the priming site or the amplified region is minimized. This suggests that the same methodology also could be applied to the mining of species-specific EST libraries, as intron amplification is a commonly cited cause of failed EST-SSR amplifications (Perez et al. 2005; Wang and Guo 2007; Kucuktas et al. 2009). The technique also ensures that sequences contaminated by undetected vectors are not used for primer design because these sequences also will not align to the outgroup genome.

EST-SSRs have been employed in a variety of studies involving marine organisms. Previously, they have been utilized for pedigree analysis (Wang et al. 2005) and incorporated into genetic linkage maps as putative "Type 1" loci (Bouza et al. 2008; Kang 2008; Kucuktas et al. 2009). In addition, studies in both aquatic and terrestrial organisms have demonstrated that EST-SSRs are a valid substitute for genomic microsatellites in a population genetics context (Ellis et al. 2006; Kim et al. 2008; Tonteri et al. 2010). The loci developed in this study demonstrated high transferability among a wide diversity of perciform fishes. As a result, markers developed with this method may be particularly suited for comparative genomics, and have the potential to become an important resource for aquaculture and genome evolution studies. Specifically, the markers designed in this study can be used to supplement the current genetic linkage map of red drum (Portnoy et al. 2010) and, once incorporated into the map, will provide a framework for comparative studies between red drum and other teleost fishes.

Ultimately, the utility of EST-SSRs as molecular markers lies in their capacity to extend functional genomic information to non-model organisms that lack species-specific genomic information. The results of this study demonstrate that a phylogenetically informed, comparative genomics approach to primer design provides a highly efficient means of exploiting EST libraries to develop EST-SSRs for organisms with no species-specific EST data. Furthermore, the process can be automated to a large extent, saving additional time and expense.

Acknowledgments We thank M. Renshaw for technical assistance in the laboratory and for providing tissues for cross-amplifications. Funding was provided by the Saltonstall-Kennedy Program (Award NA10NMF4270199) of the National Marine Fisheries Service (U.S. Department of Commerce) and by Texas AgriLife Research (Project H-6703). This paper is no. 86 in the series "Genetic Studies in Marine Fishes" and contribution no. 211 of the Center for Biosystematics and Biodiversity at Texas A&M University.

References

Adams M, Kelley J, Gocayne J, Dubnick M, Polymeropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R, Et A (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. Mol Ecol 16:907–924

Bouza C, Hermida M, Millán A, Vilas R, Vera M, Fernández C, Calaza M, Pardo BG, Martínez P (2008) Characterization of EST-derived microsatellites for gene mapping and evolutionary genomics in turbot. Anim Genet 39:666–670

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T (2009) BLAST+: architecture and applications. BMC Bioinforma 10:421

Coulibaly I, Gharbi K, Danzmann RG, Yao J, Rexroad CE (2005) Characterization and comparison of microsatellites derived from repeat-enriched libraries and expressed sequence tags. Anim Genet 36:309–315

Dawson DA, Horsburgh GJ, Küpper C, Stewart IRK, Ball AD, Durrant KL, Hansson B, Bacon IDA, Bird S, Klein Á, Krupa AP, Lee J-W, Martín-Gálvez D, Simeoni M, Smith G, Spurgin LG, Burke T (2010) New methods to identify conserved microsatellite loci and develop primer sets of high cross-species utility—as demonstrated for birds. Mol Ecol Resour 10:475–494

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99:125–132

Ellis JR, Pashley CH, Burke JM, Mccauley DE (2006) High genetic diversity in a rare and endangered sunflower as compared to a common congener. Mol Ecol 15:2345–2355

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

Housley D, Zalewski Z, Beckett S, Venta P (2006) Design factors that influence PCR amplification success of cross-species primers among 1147 mammalian primer pairs. BMC Genomics 7:253

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Ju Z, Wells MC, Martinez A, Hazlewood L, Walter RB (2005) An in silico mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, Fundulus, and Xiphophorus. In Silico Biol 5:439–463

Kang JH (2008) Genetic linkage map of olive flounder, *Paralichthys olivaceus*. Int J Biol Sci 4:143

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–510

Karlsson S, Renshaw MA, Rexroad CE III, Gold JR (2008) PCR primers for 100 microsatellites in red drum (*Sciaenops ocellatus*). Mol Ecol Resour 8:393–398

Kim KS, Ratcliffe ST, French BW, Liu L, Sappington TW (2008) Utility of EST-derived SSRs as population genetics markers in a beetle. J Hered 99:112–124

Kucuktas H, Wang S, Li P, He C, Xu P, Sha Z, Liu H, Jiang Y, Baoprasertkul P, Somridhivej B, Wang Y, Abernathy J, Guo X, Liu L, Muir W, Liu Z (2009) Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. Genetics 181:1649–1660

Li J, Li Q (2008) A set of microsatellite markers for use in the endangered sea urchin *Strongylocentrotus nudus* developed from *S. purpuratus* ESTs. Conservat Genet 9:743–745

Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings Bioinformatics 8:6–21

Perez F, Ortiz J, Zhinaula M, Gonzabay C, Calderon J, Volckaert F (2005) Development of EST-SSR markers by data mining in three species of shrimp: *Litopenaeus vannamei*, *Litopenaeus stylirostris*, and *Trachypenaeus birdy*. Mar Biotechnol 7:554–569

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

Primmer CR, Painter JN, Koskinen MT, Palo JU, Merilä J (2005) Factors affecting avian cross-species microsatellite amplification. J Avian Biol 36:348–360

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223–225

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 8:103–106

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Meth Mol Biol 132:365

Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Lowe AJ, Gibby M, Powell W (2003) How much effort is required to isolate nuclear microsatellites from plants? Mol Ecol 12:1339–1348

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehv-Slaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: perl modules for the life sciences. Genome Res 12:1611–1618

Tonteri A, Vasemägi A, Lumme J, Primmer CR (2010) Beyond MHC: signals of elevated selection pressure on Atlantic salmon (*Salmo salar*) immune-relevant loci. Mol Ecol 19:1273–1282

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4:535–538

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23:48–55

Vogiatzi E, Lagnel J, Pakaki V, Louro B, Canario AV, Reinhardt R, Kotoulas G, Magoulas A, Tsigenopoulos CS (2011) In silico mining and characterization of simple sequence repeats from gilthead sea bream (*Sparus aurata*) expressed sequence tags (EST-SSRs); PCR amplification, polymorphism evaluation and multiplexing and cross-species assays. Marine Genomics 4:83–91

Wang Y, Guo X (2007) Development and characterization of EST-SSR markers in the eastern oyster *Crassostrea virginica*. Mar Biotechnol 9:500–511

Wang H, Li F, Xiang J (2005) Polymorphic EST-SSR markers and their mode of inheritance in *Fenneropenaeus chinensis*. Aquaculture 249:107–114

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

Yu H, Li Q (2008) Exploiting EST databases for the development and characterization of EST-SSRs in the Pacific oyster (*Crassostrea gigas*). J Hered 99:208–214

Yue GH, Ho MY, Orban L, Komen J (2004) Microsatellites within genes and ESTs of common carp and their applicability in silver crucian carp. Aquaculture 234:85–98

Zhou Z-C, Zou L-L, Dong Y, He C-B, Liu W-D, Deng H, Wang L-M (2008) Characterization of 28 polymorphic microsatellites for Japanese sea urchin (*Strongylocentrotus intermedius*) via mining EST database of a related species (*S. purpuratus*). Annales Zoologici Fennici 45:4