9-14-2009

Complex Evolution of a Highly-conserved Microsatellite Locus in Several Fish Species

J.-X. Liu

Bert Ely
University of South Carolina - Columbia, ely@sc.edu

Publication Info

Postprint version. Published in *Journal of Fish Biology*, Volume 75, Issue 2, 2009, pages 442-447.

Brief Communications

Complex evolution of a highly-conserved microsatellite locus in several fish species

Jin-Xian Liu, Bert Ely*

Department of Biological Sciences, University of South Carolina, Columbia, SC, 29208, USA

*Author to whom correspondence should be addressed

Tel.: 803-777-2768; Fax: 803-777-4002; email: ely@sc.edu

Running head: Complex Microsatellite Evolution

Abstract

The evolutionary dynamics of a highly conserved microsatellite locus (*Dla 11*) was studied in several fish species. The data indicated that multiple types of compound microsatellites arose through point mutations that were sometimes followed by expansion of the derived motif. Furthermore, extensive length variation was detected among species in the regions immediately flanking the repeat region.

Key words: interruption, duplication, compound microsatellite, DNA replication slippage

Microsatellites are tandem repeats of 1-6 nucleotides scattered throughout the genomes of eukaryotes (Schlötterer, 2000; Ellegren, 2004). The vast majority of microsatellite mutations are gains or losses of repeat units, a hallmark of DNA replication slippage (Schlötterer & Tautz, 1992; Ellegren, 2000). Point mutations causing interruptions in the repeat arrays generate compound microsatellites and are critical to the evolution of many microsatellites. Within genomes at equilibrium, the distribution of microsatellite repeat lengths can be explained by a balance between length dependent replication slippage and point mutations acting toward the decay of the repetitive DNA (Ellegren, 2000). The fate of interruptions in the repeat array remains unclear. Some studies have indicated that interrupted microsatellites are less likely to mutate and that the accumulation of interruptions would lead to the degradation of microsatellites (Taylor *et al*., 1999). However, Harr *et al*. (2000) suggested that imperfect repeats could be removed through standard replication slippage, which means that point mutations in microsatellite arrays do not necessarily lead to decay, but instead they might represent a transition state during the evolution of microsatellites.

Compound microsatellite loci are abundant and represent ~10% of all microsatellites (Weber, 1990). However, the mechanisms responsible for the genesis of compound microsatellites have not been well studied. Because perfect microsatellites are much more common than the compound loci (Weber, 1990), the compound microsatellites probably arise from perfect ancestors through a process of mutation and replication slippage (Levinson & Gutman, 1987). A full understanding of these processes is essential to ensure the effective use of microsatellite loci as analytical tools.

In this study, we performed a detailed sequence analysis of a microsatellite locus (*Dla 11*) both within and among four North American temperate bass species. These species include a pair of closely related species, *Morone americana* (Gmelin) and *Morone mississippiensis* (Jordan & Eigenmann), and their relatives *Morone saxatilis* (Walbaum) and *Morone chrysops* (Rafinesque) as shown in Fig. 1 (Leclerc *et al*., 1999; Orrell *et al*., 2002).

M. saxatilis (n=11) were collected from the Congaree River in South Carolina in 1994. *M. chrysops* (n=6) were sampled from the Ohio River in Ohio in 2000. *M. americana* (n=5) were collected from North Carolina in 2000. *M. mississippiensis* (n=2) were obtained from Mississippi River in Louisiana in 1991. A specimen of *Dicentrarchus labrax* L. was collected from Marseilles, France in 1995. Genomic DNA was isolated from fin clips by DNAzol Genomic DNA Isolation Reagent (DN 127, Molecular Research Center, Cincinnati, Ohio) after a proteinase K digest.

Initial amplification of the *Dla 11* locus was performed with the primers described by Castilho & McAndrew (1998). However, indels in the 5' flanking region necessitated that the forward primer

(*Dla 11*F) be redesigned. In addition, restriction sites were added to the 5'-ends of the amplification primers to facilitate the cloning of the amplified products. Therefore, the *Dla 11* locus was amplified using primers *Dla 11*F 5' – CCCAAGCTTGGGCAAGCACACACCTCTAATGCT - 3' and *Dla 11*R 5'- CCGGAATTCCGGCGAATGCGCTACAAATCTGC - 3', with restriction-sites for *Eco*RI or *Hin*dIII (sequence underlined) near the 5'-ends. PCR was carried out in 25 µL reactions containing 1.0 µL DNA, 5.0 µL 5× Phusion HF buffer, 200 µM of each dNTP, 0.2 µM forward and reverse primers, and 0.5 U Phusion high-fidelity DNA polymerase (FINNZYMES, Espoo, Finland). The thermal cycling profile was 98 ºC for 30 s, then 30 cycles each at 98 ºC for 10 s, 53 ºC for 15 s and 72 ºC for 15 s, plus a final extension step at 72 ºC for 5 min. Amplification products were purified by GenElute PCR DNA Purification Kit (SIGMA, St. Louis, MO) and double digested with *Eco*RI and *Hin*dIII (New England Biolabs, Beverly, MA). The double digest product was purified by GenElute PCR DNA Purification Kit and ligated into pBluescript II SK + vector with T4 DNA ligase (New England Biolabs, Beverly, MA). Ligated product was transformed into *Escherichia coli* strain XL1-Blue (Stratagene, La Jolla, CA) and plated on LB agar containing 100 mg/L ampicillin, spread with 40 mL of 40 mg/mL X-gal and 40 mL of 100 mM IPTG. Four to ten white clones were randomly picked for each sample and grown in 2 mL LB medium at 37 ºC for 24 hours. The plasmid DNA was isolated by GenElute Plasmid Mini-Prep Kit (SIGMA, St. Louis, MO), and the insert was sequenced using BigDye terminator Cycle Sequencing Kit with both M13 and M13 reverse primers on a 3730 DNA analyzer (Applied Biosystems, Foster City, CA).

Nucleotide sequences were aligned and edited using DNAStar software (DNAStar, Madison, WI). The nucleotide sequences of representative alleles have been deposited in GenBank (accession numbers FJ905014-FJ905024).

The *Dla 11* locus was first identified in *D. labrax* as a perfect $(GT)_n$ repeat, with the number of repeats ranging from 6 to 27 (Castilho & McAndrew, 1998). The repeat region and most portions of 3' flanking region of *Dla 11* are highly conserved across a wide range of fish species (Fig. 2). When we conducted a BLAST search comparing the *Dla 11* nucleotide sequence (Y13159) to the five available ENSEMBL teleost whole genome assemblies, we found that the *Dla 11*sequence is found in an intron of the Transducin-like Enhancer of Split 2b gene (Aghaallaei *et al*., 2005) of *Gasteroteus aculeatus* L., *Tetraodon nigroviridis* (Marion de Procé), and *Takifugu rubripes* (Temminck & Schlegel). Based on mitochondrial genome comparisons, the divergence between *T. nigroviridis* and *G. aculeatus* occurred about 183 million years ago (Yamanoue *et al*., 2006), indicating that the *Dla 11*microsatellite has been conserved over a long evolutionary time scale. Highly conserved microsatellite loci have been found in other studies as well (FitzSimmons *et al*., 1995; Rico *et al*., 1996), challenging the general assumption that microsatellite loci are too variable to be conserved over a long timescale.

Nucleotide sequence comparisons within and among species demonstrated that the evolution of the *Dla 11* locus is quite complex. Extensive length variation was detected among species in the regions immediately flanking the repeat region (Fig. 2). This result reinforces the observation that more attention should be paid to size homoplasy that occurs both within and among species due to length variation in the microsatellite flanking regions (Brohede & Ellegren, 1999; Shao *et al*., 2005). In *M. saxatilis*, a duplication of the *Dla 11* locus must have occurred because 3 or 4 different alleles were detected for some individuals. Furthermore, two flanking region alleles were observed in *M. saxatilis*, one (FRMsax1) with short perfect repeats, and one (FRMsax2) with long and interrupted repeats

(Table I). Similarly, analysis of the *Dla 11* nucleotide sequence revealed a duplication in the *T. nigroviridis* genome, indicating that the *Dla 11* locus was duplicated in this species as well.

In addition to the flanking region changes, drastic structural evolution of the repeat region also was observed among species. It is widely assumed that microsatellite expansion could be balanced by interruption-induced stabilization (Kruglyak *et al*., 1998). However, the nucleotide sequences of the *M. saxatilis* alleles showed variation both in the number of GT repeats and in the number and distribution of interruptions, suggesting that duplication of the interruptions contributed to the length expansion. Most of the interruptions were tandemly distributed in the repeat region with a motif of $AT(GT)_4$ or $AT(GT)_2$. All of the interruptions of the GT repeat were characterized by an A in place of G. The most plausible explanation is that a G to A transition caused the initial interruption and was then propagated by DNA replication slippage. Duplication of interruptions also was suggested to explain a microsatellite locus (A113) in *Apis mellifera* L. where a TT interruption was duplicated up to three times (Estoup *et al*., 1995).

In addition to the perfect repeats found in *D. labrax* and one of the two *M. saxatilis Dla 11* loci, all of the alleles detected in *M. chrysops* were long and perfect GT repeats (Table I). The preservation of long perfect repeats in these three species indicates that interrupted and perfect repeats can co-exist over long time scales.

The structure of the *Dla 11* repeat region is also complex in the closely-related *M. americana* and *M. mississippiensis* species. A compound repeat structure $(GT)_mG_n$ was observed in both species suggesting that this compound repeat arose prior to speciation. In addition, a novel interrupted

compound allele, Mame01, was also detected in *M. Americana* (Table I). The ancestral state of *Dla 11* is probably a GT repeat, since the GT repeat is the basic repeat structure of most of the species studied to date. Therefore, the most likely molecular mechanism would be that a single T to G transversion occurred at the 3' end of the GT repeat region in a progenitor of both *M. Americana* and *M. mississippiensis*, and a mononucleotide proto-microsatellite (GGG) was generated. Replication slippage then expanded the proto-microsatellite, resulting in the compound alleles. A similar mechanism could be responsible for the repeat structure of *Dla 11* in *T. rubripes*, where a mononucleotide proto-microsatellite resulting from a G to T transversion appears to have been propagated by replication slippage (Table I). In addition, a GC repeat motif was found within the repeat region of *G. aculeatus*. In this case, the GC repeat might be generated by a T to C transition within the GT repeat array and then expanded through replication slippage (Table I). The compound status of this locus in five unrelated fish species indicates that a change from a perfect microsatellite to a compound one occurred frequently at the *Dla 11* locus.

Taken together, these results indicate that multiple types of compound microsatellites could arise from a single perfect ancestral repeat. In agreement with theoretical expectations, the compound microsatellites appear to have occurred through point mutations and followed by expansion of a derived motif within or at the end of a perfect repeat array (Levinson & Gutman, 1987). These results illustrate that length variation at microsatellite loci can arise in many ways and that identically-sized alleles may have very different evolutionary histories.

Acknowledgements

We would like to thank Jeremy Dietrick and Kurt Ash for their technical assistance and the two anonymous reviewers for their valuable comments. This work was supported in part by FISHTEC grant NOAA-NMFS RT/F-1 and NIH grant R25RR01854201.

References

Aghaallaei, N. Bajoghli, B. Walter, I. & Czerny, T. (2005). Duplicated members of the Groucho/Tle gene family in fish. *Developmental Dynamics* **234**,143-150. doi: 10.1002/dvdy.20510

Brohede, J. & Ellegren, H. (1999). Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proceedings of the Royal Society of London. Serials B: Biological Sciences* **266**, 825-833.

Castilho, R. & McAndrew, B. (1998). Two polymorphic microsatellite markers in the European seabass, *Dicentrarchus labrax* (L.). *Animal Genetics* **29**, 151-152.

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* **16**, 551-558. doi: 10.1016/S0168-9525(00)02139-9

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**, 435-445. doi: 10.1038/nrg1348

Estoup, A. Tailliez, C. Cornuet, J. M. & Solignac, M. (1995). Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). *Molecular Biology and Evolution* **12**, 1074-1084.

FitzSimmons, N. N. Moritz, C. & Moore, S. S. (1995). Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Molecular Biology and Evolution* **12**, 432-440.

Harr, B. Zangerl, B. & Schlötterer, C. (2000). Removal of microsatellite Interruptions by DNA replication slippage: phylogenetic evidence from Drosophila. *Molecular Biology and Evolution* **17**, 1001-1009.

Kruglyak, S. Durrett, R. T. Schug, M. D. & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 10774-10778.

Leclerc, G. M. Han, K. Leclerc, G. J. & Ely, B. (1999). Characterization of a highly repetitive sequence conserved among the North American *Morone* species. *Marine Biotechnology* **1**, 122-130. doi: 10.1007/PL00011759

Levinson, G. & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**, 203-221.

Orrell, T. M. Carpenter, K.E. Musick, J. A. & Graves, J. E. (2002). Phylogenetic and biogeographic analysis of the Sparidae (Perciformes: Percoidei) from cytochrome *b* sequences. *Copeia* **3**, 618-631.

Rico, C. Rico, I. & Hewitt, G. (1996). 470 million years of conservation of microsatellite loci among fish species. *Proceedings of the Royal Society of London. Serials B: Biological Sciences* **263**, 549-557.

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365-371. doi: 10.1007/s004120000089

Schlötterer, C. & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* **20**, 211-215.

Shao, Z. Lek, S. & Chang, J. (2005). Complex mutation at a microsatellite locus in sturgeons: *Acipenser sinensis, A. schrenckii, A. gueldenstaedtii* and *A. baerii*. *Journal of Applied Ichthyology* **21**, 2-6. doi: 10.1111/j.1439-0426.2004.00629.x

Taylor, J. S. Durkin, J. M. H. & Breden, F. (1999). The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. *Molecular Biology and Evolution* **16**, 567-572.

Weber, J. L. (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* **7**, 524-530.

Yamanoue, Y. Miya, M. Inoue, J. G. Matsuura, K. Nishida, M. (2006). The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. Genes and Genetic Systems. 81, 29-39.

Table I. Different alleles of repeat region among species. The allele sequence designations were defined by their organism origins. Species are designated by four-letter abbreviations of the Latin names (one from the first letter of genus name and three from the first three letters of species name, e.g. *Msax* for *M. saxatilis*). Flanking region alleles corresponding to each repeat region sequence are also shown.

Repeat region allele	Repeat sequence	Flanking region allele
Dlab01	$(GT)_{11}$	FRDlab
Dlab02[a]	$(GT)_{16}$	FRDlab
Msax1-1	$(GT)_{14}$	FRMsax1
Msax1-2	$(GT)_{20}$	FRMsax1
Msax2-1	$(GT)_{16}\mathbf{A}T(GT)_{30}$	FRMsax2
Msax2-2	$(GT)_{16}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{25}$	FRMsax2
Msax2-3	$(GT)_{31}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{21}$	FRMsax2
Msax2-4	$(GT)_{20}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{14}$	FRMsax2
Msax2-5	$(GT)_{23}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{11}$	FRMsax2
Msax2-6	$(GT)_{9}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{10}$	FRMsax2
Msax2-7	$(GT)_{11}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{11}$	FRMsax2
Msax2-8	$(GT)_{23}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{28}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{17}$	FRMsax2
Msax2-9	$(GT)_{6}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{23}$	FRMsax2
Msax2-10	$(GT)_{19}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{23}$	FRMsax2
Msax2-11	$(GT)_{24}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{11}$	FRMsax2
Msax2-12	$(GT)_{19}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{2}\mathbf{A}T(GT)_{20}$	FRMsax2
Msax2-13	$(GT)_{29}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{4}\mathbf{A}T(GT)_{7}\mathbf{A}T(GT)_{12}$	FRMsax2
Mchr	$(GT)_{40\sim69}$	FRMchr
Mame01	$(GT)_{8}(G)_{9}T(G)_{6}T(G)_{5}T$	FRMame1
Mame02	$(GT)_{15}(G)_{21}$	FRMame2
Mame03	$(GT)_{13}(G)_{16}$	FRMame1
Mame04	$(GT)_{7}(G)_{12}$	FRMame1
Mmis01	$(GT)_{8}(G)_{19}$	FRMmis
Mmis02	$(GT)_{6}(G)_{25}$	FRMmis
Tnig	$(GT)_{6}CC(GT)_{16}$	FRTnig
Trub	$(GT)_{3}(T)_{7}(GT)_{2}GA(GT)_{3}GA(GT)_{7}$	FRTrub
Gauc	$(GT)_{3}GC(GT)_{3}(GC)_{3}(GT)_{22}$	FRGauc

[a]data from Castilho & McAndrew (1998)

Figures

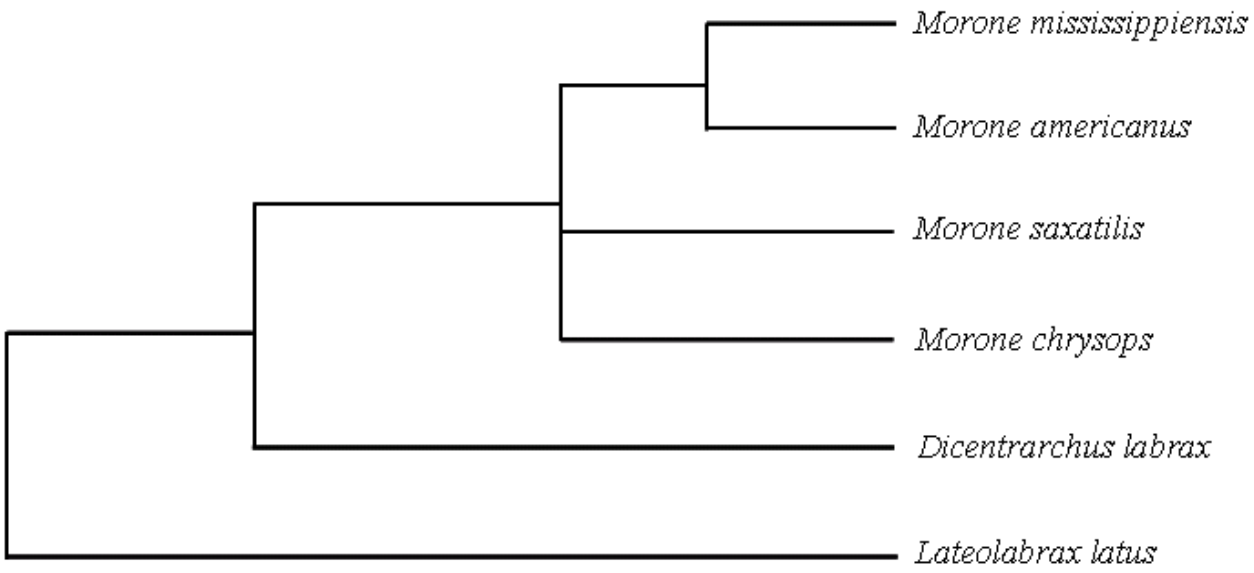

Figure 1. Phylogeny of *Morone* species and *D. labrax*. This figure depicts a strict consensus tree of two equally parsimonious trees inferred from the weighted cytochrome b nucleotide data (Orrell *et al*., 2002).

```
       {  5' flanking region     }              {              3' flanking region                        }
FRDlab  CAAGCACACACCTCTAATGCTTCCATGCT-- [repeat region] TCAGTATAGTGCAGCACTGACAGAGCTGTCTGAAGGCTGGCAGGCTGCAGATTTGTAGCGCATTCG
FRMsax1 ..............................CC [repeat region] --.....................................................................
FRMsax2 ..............................CC [repeat region] --G....................................................................
FRMchr  ......................-------- [repeat region] .....................................................................
FRMame1 ......................TTT------ [repeat region] -----....................................................................
FRMame2 ......................TTT------ [repeat region] ----A....................................................................
FRMmis  ......................TTT------ [repeat region] ----.....................................................................
FRGacu                                  [repeat region] ----CGC...AG.....................................................................
FRTnig                                  [repeat region] ----------.G.....................................A.....A
FRTrub                                  [repeat region] ----------AT.....................................A
```

Figure 2. Alignments of all the unique flanking region alleles of *Dla 11*. Dots indicate identical nucleotides, and dashes indicate indels. The sequences in the primer regions were underlined. Nomenclature of the flanking region alleles of different species were designated as in Table I with FR at the beginning. The 3' flanking region sequences of *G. aculeatus, T. nigroviridis,and T. rubripes* were from published genomic sequence data, and the 5' flanking regions were not shown because of low sequence homology to the *Morone* species.