Letter to the Editor

Allelic Diversity in Alligator Microsatellite Loci Is Negatively Correlated with GC Content of Flanking Sequences and Evolutionary Conservation of PCR Amplifiability

Travis C. Glenn,† Wolfgang Stephan,† Herbert C. Dessauer,‡ and Michael J. Braun**

*Laboratory of Molecular Systematics, Smithsonian Institution; †Department of Zoology, University of Maryland; and ‡Department of Biochemistry, Louisiana State University

Microsatellites are short tandem arrays of simple DNA sequences, having a basic repeat unit of 1-5 bp (Tautz and Schlbtterer 1994). Microsatellite loci are abundant in most complex genomes (Tautz and Schlötterer 1994) and are usually highly polymorphic when the basic repeat unit is duplicated more than 10 times (Weber 1990). Alleles at a locus differ in length due to variation in the number of repeats, and slipped strand mispairing is considered to be the major mechanism by which length variants (new alleles) are generated (Levinson and Gutman 1987; Strand et al. 1993). Variation at microsatellite loci can be assayed by PCR amplification using primers complementary to unique sequences flanking specific repetitive arrays, followed by electrophoretic sizing of the PCR products (Tautz 1989). Because microsatellites are highly polymorphic and amenable to PCR, multiplex PCR, and automated scoring, they often are the molecular genetic marker of choice (e.g., Bruford and Wayne 1993).

Although microsatellites occur in all eukaryotes (Tautz and Schlötterer 1994), their distribution and overall frequency varies widely (e.g., Langercrantz, Ellegren, and Andersson 1993). Because primer development can be time-consuming and costly, many investigators attempt to use primers developed in one species to amplify orthologous loci in closely related species (e.g., Moore et al. 1991). Some of these studies show that this strategy can be quite successful, even in distantly related species (e.g., FitzSimmons, Moritz, and Moore 1995). Most such cross-species comparisons, however, have tested either a small number of primer pairs (e.g., FitzSimmons, Moritz, and Moore 1995) or a small number of taxa (e.g., Moore et al. 1991). Additionally, only loci that are polymorphic in the species from which they are developed are used in most studies. The selection and use of such loci biases cross-taxa comparisons (Ellegren, Primmer, and Sheldon 1995). In general, primers that exhibit low levels of variation, as well as those that yield complex or unclear patterns, are discarded without much investigation or hesitation. Thus, it is not known how a reasonably large random sample of microsatellite sequences evolves within a group of taxa.

We have developed primers to amplify microsatellite loci from American alligators *(Alligator mississippiensis).* Details of the lab work will be published elsewhere; protocols may be downloaded from the Laboratory of Molecular Systematics' Internet server by anonymous FTP (onyx.si.edu/protocols/MsatManV6.rtf). During development, loci were maintained in the study if: (1) they contained at least 10 AC or AG repeats and (2) they regularly yielded clear amplification patterns consistent with a single locus when sized on agarose or acrylamide. Below, we present an analysis of the first 14 loci which met these criteria.

Amplification was attempted for all 14 loci using DNA from at least one individual representing each of the eight extant crocodilian genera *(Alligator sinensis, Caiman crocodilus, Melanosuchus niger, Paleosuchus palpebrosus, Tomistoma schlegelli, Gavialis gangieticus, Osteolamus tetraspis, Crocodylus cataphractus, C. niloticus,* and C. *rhombifer).* We used amplification conditions optimized for alligators (i.e., the most stringent conditions that yielded products in alligators) to test all taxa. For each locus, the number of crocodilian genera that gave amplification was determined (summarized in table 1). Cross-taxa amplification products were sized on 2% agarose gels containing ethidium bromide, and scored as: present with similar intensity as the alligator amplicons (1), present but with much less intensity than the alligator amplicons (0.5), or not present (0). For the genus *Crocodylus,* in which three species were amplified, the total score for the genus was normalized to one (e.g., if the three species scored 1, 0.5, and 0, then the score for the genus was $[(1 + 0.5 + 0)/3]=0.5$).

Genotypes for 28 alligators (14 from Rockefeller Refuge, La., and 14 from Everglades National Park, Fla.) were determined for all 14 loci using standard methods (unpublished data, results summarized in table 1). Genotypes for an additional 15 alligators (5 from Rockefeller Refuge, La., and 10 from Terrebonne Parish, La.) were determined for the 11 polymorphic loci (table 1). Finally, one individual from each of four taxa (*Alligator sinensis, Paleosuchus palpebrosus, Tomistoma schlegelli,* and *Crocodylus cataphractus* or C. *porosus)* was amplified and scored with the alligators on 6% Long Ranger (acrylamide) gels.

All 14 loci gave clear patterns consistent with single-locus products in alligators. Patterns were also clear, when present, in all other taxa tested. Additionally, all products scored were similar in size to the amplicons from alligators. Relaxation of the stringency of PCR conditions resulted in additional high molecular weight PCR products and/or the reduction of products of the expected size. Thus, the high stringency of PCR seemed crucial for clear results.

While compiling our results, we noted several relationships. To assess the significance of these relationships, we calculated Spearman's rank correlation (rho)

Key words: *Alligator mississippiensis,* Crocodilia, GC content, interspecific priming, polymorphisms, microsatellites, SSR, STR.

Address for correspondence and reprints: Travis C. Glenn, MRC 534, Smithsonian Institution, Washington, D.C. 20560. E-mail: glenn@onyx.si.edu.

Mol. Biol. Evol. 13(8):1151–1154. 19%
 ISSN: 0737-4038

Table 1
Characteristics of Microsatellite Loci Amplified from American Alligators

Locus	Alleles[a]	H_s[b]	Var.[c]	Genera[d]	Temp.[e]	%GC[f]	Repeat Sequences[g]
AC-19	1	0.000	0.000	5.50	60	48.3	$(AC)_1(CT)_{11}$
AC-12	1	0.000	0.000	7.00	47	42.4	$(AC)_{13}(AT)_9(AT)_7$
AC-1	1	0.000	0.000	8.00	50	48.1	$(AC)_1$
AG-1	1	0.000	0.000	8.00	50	43.8	(AC)„
AC-13	4 (4)	0.539	13.3	4.83	60	52.4	$(AC)_{>16}$
AC-3	5 (5)	0.665	7.03	7.00	50	38.2	(AC)„
AC-16	5 (6)	0.490	28.2	6.00	60	36.4	(AC)„
AC-15	7 (8)	0.617	23.3	2.00	60	39.0	(AC)„
AC-18	7 (8)	0.718	72.6	4.67	60	37.2	(AC)„
AC-6	7 (9)	0.612	28.8	3.00	55	33.0	(AC)„
AC-20	8 (8)	0.756	120	6.83	55	33.2	(AC)„
AC-8	10(11)	0.750	26.2	4.33	55	38.8	(AC)>„
AC-17	14 (17)	0.771	168	4.83	55	37.7	$(AC)_{22}(AT)_7(ATGT)_{12}(AGAT)_9$
AG-2	16 (17)	0.817	158	2.00	57	31.1	$(AG)_{20}(AT)_{20}$

[a] Number of alleles detected among 28 or (43) alligators from Louisiana and Florida.
[b] Average within-group expected heterozygosity (Nei and Chesser 1983) among 28 alligators from the Rockefeller and Everglades populations. Calculated using GeneStrut (Constantine, Hobb, and Lymbery 1994).
[c] Variance of allele sizes among 28 alligators from Louisiana and Florida.
[d] Number of crocodilian genera amplified; method of calculation is given in the text.
[e] Annealing temperature used in PCR (°C).
[f] Percentage of G plus C bases in the sequence of the cloned allele excluding the repeat sequences.
[g] Sequence of the longest perfect run of repeats and all adjacent repeats within the cloned allele.

using JMP version 3.1 for the Macintosh (SAS Institute). The inclusion of the 15 alligators genotyped for only the polymorphic loci resulted in *P* values within 0.005 of correlations where only the 28 alligators genotyped for all loci were considered. Thus, those 15 alligators are dropped from further discussion of results.

Three expected relationships were observed. First, as originally noted by Weber (1990), the number of alleles at a locus is correlated with the number of repeats in the longest perfect run of repeats in the sequenced allele (table 1, rho = 0.7023, *P* = 0.0109). Second, the likelihood that a taxon will amplify for a particular locus is related to the taxon's phylogenetic distance from alligators. For example, the primers tested are most efficient when amplifying other species in the Alligatoridae (77% of species/locus combinations) or Gavialidae (*Tomistoma* and *Gavialis*; 80% of species/locus combinations), but are much less efficient when amplifying Crocodylidae (38% of species/locus combinations), the family most distantly related to the alligator (Densmore 1983). Third, the number of genera amplified is significantly negatively correlated with the annealing temperature used for the PCR reactions (rho = -0.6162, *P* = 0.0189).

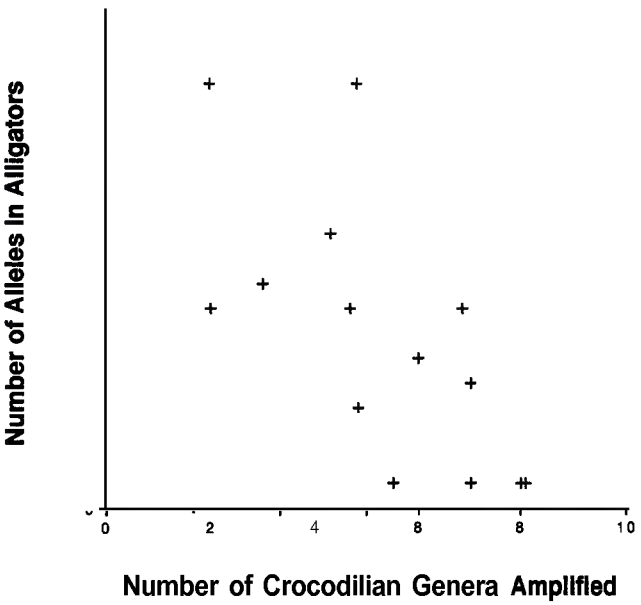

FIG. 1 .-Number of alleles detected in 43 alligators in relation to the number of genera of crocodilians for which each locus could be amplified. Each point represents one microsatellite locus amplified with one primer pair (see text for details).

Two unexpected correlations were also observed. First, the GC content of the sequences flanking the microsatellites is negatively correlated with the number of alleles in alligators (Table 1; rho = -0.7420, *P* = *0.0024).* The GC content is also significantly correlated with other measures of allelic diversity, such as the average within-group expected heterozygosity (H_s, table 1; rho = -0.6823, *P* = *0.0072)* and the variance in allele sizes (table 1; rho = -0.8112, *P* = 0.0004). GC content is not correlated with the number of genera amplified or with the annealing temperature used in PCR. Second, the number of alleles per locus in alligators was negatively correlated with the number of crocodilian genera amplified (fig. 1; rho = -0.7026, *P* = 0.0051). The correlation remains significant ($P < 0.05$) if each taxon is weighted equally (i.e., *Crocodylus* is not normalized to one) and/or if 0.5 scores are given values of 0 or 1. The number of genera amplified is also significantly correlated with the average within-group expected heterozygosity (H_s, table 1; rho = -0.5860, *P* = 0.0277) and the variance in allele sizes (table 1; rho = -0.6060, *P* = 0.0216).

Because microsatellites occur in all isochore classes with similar frequency (Duret, Mouchiroud, and Gau-

tier 1995), it is reasonable to assume that a random sample of microsatellites will contain representatives of all classes. The percentage of GC is lower in pseudogenes than in functional genes (e.g., Gillespie 1991, p. 90), and varies with substitution rates (Wolfe, Sharp, and Li 1989). Thus, the correlation of GC content with allelic diversity may derive from the chromosomal position of the microsatellite and be related to the likelihood that it is within or near functional genes or experiences differing mutational parameters. Models that incorporate correlations of GC content with mutation rate (Gu and Li 1994) and mismatch repair (Eyre-Walker 1994) have been proposed. The correlation of GC content with allelic diversity also implies that GC content can be used as a criterion for the selection of clones with the highest likelihood of polymorphism.

We considered the possibility that the observed correlation between allelic diversity and the number of genera amplified might be a trivial corollary of the stringency of PCR annealing temperatures, because these were also correlated with the number of genera amplified. However, annealing temperature is not correlated with alleles per locus. Additionally, a multiple-regression analysis using genera amplified and annealing temperature to predict the number of alleles does not reveal a relationship between annealing temperature and alleles. There is also no correlation between the melting temperature of the primers (the higher of the pair, lower of the pair, or average of the pair) and the annealing temperature used for amplifications. Thus, the correlation of the number of genera amplified with allelic diversity (fig. 1) is not due to the correlation of the annealing temperature used in PCR with the number of genera amplified.

If one assumes that allelic diversity reflects mutations within the repeats and that amplifiability reflects mutations accumulated at primer binding sequences, then there is a correlation between variation in the number of repeat copies and variation in the DNA flanking the repeats (where the primers anneal). Furthermore, because changes in copy number are probably due to replication slippage (Levinson and Gutman 1987; Strand et al. 1993), and the changes of flanking DNA are due to insertions, deletions, or nucleotide substitutions (collectively called nucleotide mutations here), our data suggest a correlation between these two classes of sequence variation.

There are at least two sorts of mechanisms which might produce the observed negative correlation between microsatellite allelic diversity and evolutionary conservation of flanking sequences. The first assumes that the various alleles at each locus are selectively equivalent (neutral). The second supposes that there are functional or selective constraints on some loci which limit polymorphism and sequence divergence (cf. Kunkel 1993). The various mechanisms briefly discussed below are not mutually exclusive; several or all might operate at once to produce the observed correlations.

In the first case, it is necessary to postulate a direct relationship between the generation (and maintenance) of length variants in the repetitive element and the rate of evolution of flanking sequences. Such a relationship would be obtained if either (1) mutation rates in repeats and flanks are correlated or (2) repair rates in repeats and flanks are correlated. Microsatellites are often found in stretches of DNA that contain many short repetitive elements which may lower the fidelity of DNA replication through strand-slippage events. Alternatively, the biased nucleotide composition of microsatellite repeats may result in temporary reductions in the local concentration of some nucleotides during replication, shortening processivity and increasing polymerase error rates. Either of these mutational mechanisms might conceivably affect mutation rates in both repeats and flanks. Our observations are also consistent with mechanisms which invoke incomplete or error-prone DNA repair. The rates of both slipped-strand and single nucleotide mutations are known to be dependent on DNA repair processes (Strand et al. 1993), and microsatellite instability is increased when mismatch repair is defective (Drummond et al. 1995).

Alternatively, functional constraints on some microsatellite loci may limit both polymorphism in repeat copy number and sequence evolution in flanking regions, producing the observed correlation. If some microsatellites lie in protein-coding sequences, serve as binding sites for other molecules, or are sufficiently close to such potentially selected sites, then they might be expected to exhibit reduced variability due to functional constraints or hitchhiking effects. If stabilizing selection is occurring on a particular locus throughout a group of taxa, then one would expect to find fragments of about the same (optimal) size in all taxa (Walsh 1987). Indeed, for locus AC-12 the fragment sizes were identical in all four additional taxa investigated on gels with single-nucleotide resolution. However, for other monomorphic loci, there was variation in fragment sizes among the additional taxa.

Finally, it is also important to note that all four monomorphic alligator loci have 10–15 uninterrupted repeats, whereas all polymorphic loci have 16 or more uninterrupted repeats in the cloned allele (table 1). This suggests that the minimum number of repeats necessary for variability of AC and AG microsatellites in alligators is greater than the number necessary for other vertebrates investigated thus far, where about 90% of loci with more than 10 repeats are polymorphic. Observations of only 5% polymorphism among AC microsatellites in *Arabidopsis* (Bell and Ecker 1994) suggest that very different mutational mechanisms or selective constraints may be in operation for microsatellite loci in various taxa.

Acknowledgments

We wish to acknowledge the assistance of D. Albright, R. Atkinson, R. Brumfield, K. Coulon, L. Densmore, R. Elsey, C. Huddleston, T. Joanen, J. Kushlan, L. McNease, R. Menzies, J. Sakamoto, K. Winker, the Audubon Zoological Garden, the Louisiana Land and Exploration Co., and the LSU Museum of Zoology. We thank R. Harrison and two reviewers for suggesting sig-

nificant improvements in the manuscript. This project was funded by the Smithsonian's Lab of Molecular Systematics and the Louisiana Department of Wildlife and Fisheries.

LITERATURE CITED

BELL, C. J., and J. R. ECKER. 1994. Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*. Genomics **19**:137–144.

BRUFORD, M. W., and R. K. WAYNE. 1993. Microsatellites and their application to population genetic studies. Curr. Opin. Genet. Dev. **3**:939–943.

CONSTANTWE, C. C., R. P. HOBBS, and A. J. LYMBERY. 1994. FORTRAN programs for analyzing population structure from multilocus genotypic data. J. Hered. **85**:336–337.

DENSMORE, L. D. III. 1983. Biochemical and immunological systematics of the order Crocodilia. Pp. 397-465 *In* M. K. HECHT, B. WALLACE, and G. T. PRANCE, eds. Evolutionary biology. Vol. 15. Plenum, New York.

DRUMMOND, J. T., G. LI, M. J. LONGLEY, and F? MODRICH. 1995. Isolation of an hMSH20-p160 heterodimer that restores DNA mismatch repair to tumor cells. Science 268: 1909-1912.

DURET, L., D. MOUCHIROUD, and C. GAUTIER. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40**:308–317.

ELLEGREN H., C. R. PRIMMER, and B. C. SHELDON. 1995. Microsatellite 'evolution': directionality or bias? Nat. Genet. **11**:360–362.

EYRE-WALKER, A. 1994. DNA mismatch repair and synonymous codon evolution in mammals. Mol. Biol. Evol. 11: 88-98.

FITZSIMMONS, N. N., C. MORITZ, and S. S. MOORE. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Mol. Biol. Evol. 12: 432–440.

GILLESPIE, J. E. 1991. The causes of molecular evolution. Oxford University Press, New York.

Gu, X., and W.-H. LI. 1994. A model for the correlation of mutation rate with GC content and the origin of GC-rich isochores. J. Mol. Evol. **38**:468–475.

KUNKEL, T. A. 1993. Slippery DNA and diseases. Nature 365: 207-208.

LANGERCRANTZ, U., H. ELLEGREN, and L. ANDERSSON. 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. **21**:1111–1115.

LEVINSON, G., and G. A. GUTMAN. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4**:203–221.

MOORE, S. S., L. L. SARGEANT, T. J. KING, J. S. MATTICK, M. GEORGES, and J. S. HETZEL. 1991. The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. Genomics **10**:654–660.

NEI, M., and R. K. CHESSER. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. **47**:253–259.

STRAND, M. T., A. PROLLA, R. M. LISKAY, and T. D. PETES. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature **365**:274–276.

TAUTZ, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17**:6463–647 1.

TAUTZ, D., and C. SCHLÖTTERER. 1994. Simple sequences. Curr. Opin. Genet. Dev. **4**:832–837.

WALSH, J. B. 1987. Persistence of tandem arrays: implications for satellite and simple sequence DNAs. Genetics 115:553-567.

WEBER, J. L. 1990. Informativeness of human $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphisms. Genomics **7**:524–530.

WOLFE, K. H., P. M. SHARP, and W.-H. LI. 1989. Mutation rates differ among regions of the mammalian genome. Nature **337**:283–285.

RICHARD G. HARRISON, reviewing editor

Accepted June 7, 1996