MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) doi: 10.1111/1755-0998.12000

Thousands of microsatellite loci from the venomous coralsnake *Micrurus fulvius* and variability of select loci across populations and related species

TODD A. CASTOE,*§[1] JEFFREY W. STREICHER,†[1] JESSE M. MEIK,† MATTHEW J. INGRASCI,† ALEXANDER W. POOLE,* A. P. JASON DE KONING,* JONATHAN A. CAMPBELL,† CHRISTOPHER L. PARKINSON,‡ ERIC N. SMITH† and DAVID D. POLLOCK*

*Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, 12801 17th Avenue, Aurora, CO 80045, USA, †Department of Biology & Amphibian and Reptile Diversity Research Center, The University of Texas at Arlington, 701 S. Nedderman Dr, Arlington, TX 76019 USA, ‡Department of Biology, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816 USA

Abstract

Studies of population genetics increasingly use next-generation DNA sequencing to identify microsatellite loci in nonmodel organisms. There are, however, relatively few studies that validate the feasibility of transitioning from marker development to experimental application across populations and species. North American coralsnakes of the *Micrurus fulvius* species complex occur in the United States and Mexico, and little is known about their population structure and phylogenetic relationships. This absence of information and population genetics markers is particularly concerning because they are highly venomous and have important implications on human health. To alleviate this problem in coralsnakes, we investigated the feasibility of using 454 shotgun sequences for microsatellite marker development. First, a genomic shotgun library from a single individual was sequenced (approximately 7.74 megabases; 26 831 reads) to identify potentially amplifiable microsatellite loci (PALs). We then hierarchically sampled 76 individuals from throughout the geographic distribution of the species complex and examined whether PALs were amplifiable and polymorphic. Approximately half of the loci tested were readily amplifiable from all individuals, and 80% of the loci tested for variation were variable and thus informative as population genetic markers. To evaluate the repetitive landscape characteristics across multiple snakes, we also compared microsatellite content between the coralsnake and two other previously sampled snakes, the venomous copperhead (*Agkistrodon contortrix*) and Burmese python (*Python molurus*).

Keywords: Elapidae, high-throughput marker identification, Seq-to-SSR approach, simple sequence repeats, snake genomics

Received 11 May 2012; revision received 18 July 2012; accepted 28 July 2012

Introduction

The use of next-generation DNA sequencing to identify large numbers of microsatellite loci, or simple sequence repeats (SSRs), is increasingly employed by researchers in the fields of ecology, conservation biology and population genetics (Gardner *et al.* 2011; Guichoux *et al.* 2011). This type of microsatellite development is rapidly becoming more affordable (Castoe *et al.* 2009, 2012; Santana *et al.* 2009; Jennings *et al.* 2011) and provides a means for researchers working with modest budgets to identify massive numbers of microsatellite loci from nonmodel organisms. These approaches usually identify SSR loci from a single individual by sequencing simple genomic shotgun libraries (Castoe *et al.* 2009, 2012), transcriptomes (Mikheyev *et al.* 2010) or genomic libraries enriched for specific microsatellites through hybridization (Santana *et al.* 2009). As costs of sequencing have gone down, however, identification of microsatellite loci from unassembled reads of genomic shotgun libraries (Seq-to-SSR) appears to be the most effective and economical for most applications (Castoe *et al.* 2012). Validation experiments to examine the utility of such

Correspondence: David D. Pollock, Fax: 303-724-3215; E-mail: David.Pollock@ucdenver.edu

§Present address: Department of Biology The University of Texas at Arlington 701 S. Nedderman Dr Arlington TX 76019 USA.

[1]The first two authors contributed equally.

microsatellites based on successful amplification and variation at the intraspecific (e.g. Delmas *et al.* 2011) and interspecific levels (e.g. Mikheyev *et al.* 2010) have been informative and encouraging, but have only been conducted on a handful of taxa (e.g. insects and plants).

In this study, we focus on SSR development in coralsnakes, motivated by our long-term interests in the population genetics and adaptive evolution of natural venomous snake populations. The colourful coralsnakes of the genus *Micrurus* belong to the family Elapidae, a family of venomous snakes that also includes sea snakes, mambas, cobras and kraits (Campbell & Lamar 2004). The genus includes over 70 species that are largely confined to the New World tropics. One exception is the northernmost group, the *Micrurus fulvius* species complex, which enters subtropical zones of the southeastern United States. Currently, this group is thought to contain four species: *M. bernadi*, *M. fulvius*, *M. tamaulipensis* and *M. tener* (Lavin-Murcio & Dixon 2004). Some authors further subdivide *M. tener* into four subspecies: *tener*, *fitzengeri*, *maculatus* and *microgalbineus* (Ersnt & Ersnt 2003). Species of the *M. fulvius* complex inhabit the southeastern United States from North Carolina and Florida, west along the Gulf Coast to central Texas and south to southern Veracruz, Mexico. From Veracruz, the distribution of one species (*M. bernadi*) extends inland through the highlands of Central Mexico to the state of Morelos. Across this large geographic range, the *M. fulvius* complex is somewhat variable in colour pattern but otherwise morphologically invariant (Campbell & Lamar 2004). Population genetic structure has yet to be thoroughly examined in any species of this complex, despite the knowledge that venom composition varies both geographically and phylogenetically in these taxa (Sánchez *et al.* 2008; Salazar *et al.* 2011). Thus, we are motivated to understand population genetic structure across this species complex, and its relevance to taxonomy, historical biogeography and envenomation.

Our goal in this study was to develop an approach for rapidly and inexpensively identifying informative, polymorphic microsatellite loci for a previously unstudied species by leveraging a hierarchical sampling design that incorporated high-throughput sequencing, screens for amplification and variation and then scoring of the most promising loci. We used 454 high-throughput sequencing to obtain a moderate amount of sequence (approximately 7.74 megabase pairs, or Mbp; 26 831 reads) from a genomic library made from an individual coralsnake (*Micrurus fulvius*). We identified thousands of SSR loci with flanking sequences suitable for PCR primers, and from these, we selected a set of dinucleotide and trinucleotide repeat loci to screen. Evidence of successful PCR amplification and allelic size variation across a subset of samples was obtained using an Agilent Bioanalyzer and inexpensive nonlabelled, locus-specific primer sets. In the final phase, we used the information from this screen to test (i) cross-species amplification, (ii) marker polymorphism and (iii) the utility for population genetics studies of selected microsatellite loci using dye-labelled primer sets scored *en masse* on an automated sequencer. Additionally, we compare the genomic simple sequence repeat landscape of the coralsnake to two other snake species previously sampled.

Materials and methods

DNA isolation and microsatellite identification

A single specimen of *Micrurus fulvius* from Putnam County, Florida, USA (University of Florida voucher number UF72716, field number SBH266053), was used as the source of tissue for the construction of a whole-genome shotgun library. DNA was extracted from liver tissue that had been stored at −80 °C immediately following tissue sampling. We used the same protocols described in the study by Castoe *et al.* (2009) for digesting tissues, extracting DNA, shotgun library preparation and 454 sequencing using 454 FLX Titanium reagents and protocols.

We used the program *PAL_FINDER_v0.02.04* (Castoe *et al.* 2012) to extract reads that contained perfect dinucleotide (2mer), trinucleotide (3mer), tetranucleotide (4mer), pentanucleotide (5mer) and hexanucleotide (6mer) tandem SSRs (available from http://www.snakegenomics.org/ToddCastoe/Software.html and http://sourceforge.net/projects/palfinder/). This current version of the program is appropriate for both Illumina and 454 data, while its predecessor program *PRIMERDESIGNER* (Castoe *et al.* 2009) only works on 454 data. Following the default settings described in the study by Castoe *et al.* (2009, 2012), raw reads were identified as containing SSR loci if they contained simple repeats of at least 12 bp in length for 2–4mers (e.g. 6 tandem repeats for 2mers) and at least 3 tandem repeats for 5mers or 6mers.

Where possible, suitable PCR primer sequences were designed in regions flanking SSR loci using the default settings of *PALFINDER*, which uses *PRIMER3* (Rozen *et al.* 2000) for the core primer design process; these are referred to as PALs. Because primer sequences are unlikely to work effectively to amplify a single locus if they occur multiple times in the genome, *PALFINDER* uses the entire genomic shotgun data set as a means of identifying primer sequences that might be highly repeated, and thus less likely to amplify their intended target locus exclusively. Specifically, for each designed primer, its observed frequency in the entire shotgun genomic sequence collected is counted, as is the frequency of pairs

of primers together in any one sequence read (see Castoe *et al.* 2009, 2012).

Selection of candidate potentially amplifiable loci

For amplification and validation of PALS, we targeted loci with 2mer and 3mer repeats because they are expected to have higher mutation rate compared to loci with longer tandem repeats (Del Giudice *et al.* 2001). We further narrowed our choice of candidate loci according to the following four criteria: exclusion of (i) repeats that contained repeats of the same nucleotide (e.g. TTC or AGG), (ii) repeats with similar flow orders to the 454 pyrosequencing chemistry (TACG), (iii) repeats known to be associated with transposable elements (i.e. Castoe *et al.* 2011) and (iv) PALs with priming sites that occur frequently in the genomic sequence sample. This stringent filtering resulted in 86 primer sets (62 for 2mer loci and 24 for 3mer loci). Hereafter, we refer to these markers as target potentially amplifiable loci (tPALs). From tPALs, we selected a subset of 40 (31.2mers and 9.3mers) for PCR screening in multiple individual samples representing the *M. fulvius* complex.

Additional geographic sampling

We acquired 76 snake tissue samples from throughout the geographic range of the *Micrurus fulvius* complex (Fig. 1; Table S1), including representatives of all species and subspecies. For comparative purposes, we also included a distantly related species of coralsnake, *M. browni*. Muscle tissue, liver tissue or shed skins were taken and preserved in either an SDS-based lysis buffer solution (Burbrink *et al.* 2009) or 95% ethanol. Some of the shed skins were stored dry, and so prior to isolation, we incubated these samples in lysis buffer at 56 °C for 3 days. Genomic DNA was isolated from these samples using standard protocols for the DNeasy kit (Qiagen, Valencia, CA, USA).

Rapid stepwise microsatellite screening

To determine whether the tPALs would amplify across individuals, and to identify polymorphic loci, we first made comparisons using four individuals: an *Micrurus fulvius* from North Carolina, USA, and three *M. tener* from various localities in Texas, USA. These four samples were amplified in 20-μL reaction volume that included 12 μL of Green Master Mix (Promega, Madison, WI, USA), 6 μL of water, 2 μL of primer and 2 μL of DNA template. We used a standard thermal cycling profile of 5-min initial 95 °C denaturation followed by 30 cycles of 95 °C for 30 s, 50 °C for 30 s and 60 °C for 1 min and a final extension cycle at 60 °C for 30 min.

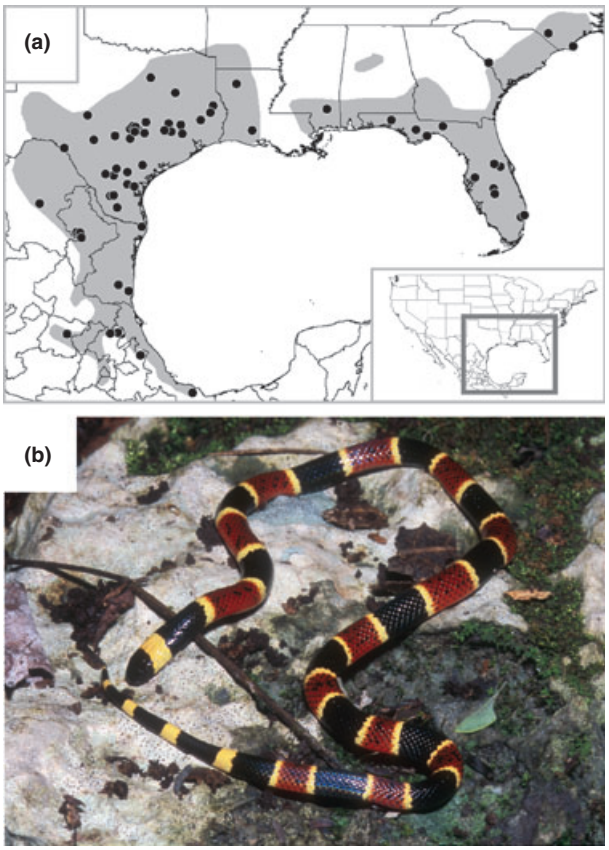

Fig. 1 Geographic range and sampling of the *Micrurus fulvius* species complex. (a). Geographic sampling of *Micrurus fulvius* complex snakes used in the present study (shaded area indicates presumed composite range of the complex) and (b) *M. fulvius* (ENS 10801) from Seminole County, Florida, USA.

Loci among the initial 40 that amplified consistently across all four individuals were further screened using a Bioanalyzer 2100 Expert (Agilent, Santa Clara, CA, USA) to identify size polymorphism. To estimate whether these loci were putatively polymorphic, we visually compared allele fragment sizes from the four individual snakes by overlaying Bioanalyzer electropherograms using the Agilent 2100 Expert software package.

Putatively polymorphic PALs were screened across all 76 individuals in our sample using PCR. This screening used the same protocols described above, followed by visualization of products on agarose gels. Once a particular locus was found to be amplifiable across the 76-individual data set, we scored fragments for length (in base pairs) using a Qiagen multiplex PCR kit and 5′ fluorescent labels on one of the two primers for each locus (Integrated DNA Technologies, Coralville, IA, USA). Fragments were separated on a 3130× genetic analyzer (ABI, California, USA) using ABI GeneScan-500 LIZ as a size standard. Allelic sizes were scored using *GENEMAPPER* version 3.5 (ABI) and manually verified.

The microsatellite allele variation data we generated were evaluated using the program *GenAlEx* 6.41 (Peakall & Smouse 2006) to perform a Mantel test for isolation by geographic distance among 50 individuals (including representatives from all four species) using codominant genotypic distances calculated from five microsatellite loci. We investigated whether the variable PALs contained sufficient variation to differentiate between the closely related species *M. fulvius* and *M. tener* that are separated by the Mississippi River, and were considered subspecies of *M. fulvius* until recently (Campbell & Lamar 2004). This hypothesis was evaluated using the program *STRUCTURE* v2.3.1 (Pritchard *et al.* 2000) to test for evidence of population structure (e.g. two distinct populations). Between one and seven populations (K) were assumed using a burn-in of 50 000 iterations followed by 500 000 iterations of the Markov chain. Simulations for each *K* value were run independently at least three times to evaluate reproducibility of runs.

Results

Microsatellite loci identified in Micrurus fulvius

A total of 26 831 reads were obtained from the 454 shotgun library of *Micrurus fulvius*, with a mean read length of 288 and 7 735 311 bp of total sequence. We identified 3840 reads (14.3% of all reads) that contained microsatellite loci meeting our length criteria (Fig. 2), which were deposited in GenBank under the accession numbers JX036543-JX040307 and are also available as supplementary material online (Table S1).

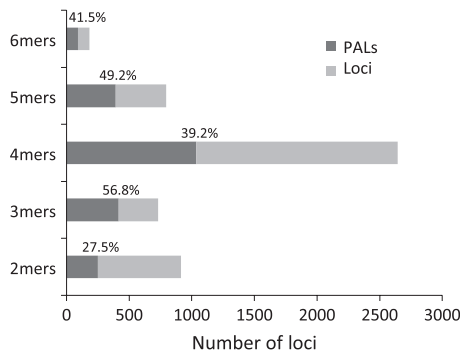

Fig. 2 Microsatellite loci frequencies. Numbers of identified microsatellite repeat loci (light shading) and the number of loci with suitable flanking PCR primer sites (referred to as 'potentially amplifiable loci', PALs; dark shading), in 26 831 reads randomly sampled from the genome of an individual *Micrurus fulvius* by 454 sequencing. The per cent of all microsatellite loci for which flanking primers were successfully designed are indicated (per repeat length class).

In total, 7.0% of all reads and 49% of all microsatellite-containing reads contained a potentially amplifiable locus (i.e. microsatellites flanked by suitable PCR priming sites, or PALs). Some reads contain multiple microsatellite loci, and *PALFINDER* designs flanking primers to incorporate the longest microsatellite locus in such multi-microsatellite reads. A tab-delimited file with information for the 1871 PALs, including primer sequences, locus ID, and information about the tandem repeats observed per locus is included as supplementary information (Table S2). Broken down by repeat motif lengths, and as shown in Fig. 2, there were 913 2mer loci (including 251 PALs), 731 3mer loci (415 PALs), 2645 4mer loci (1036 PALs), 794 5mer loci (391 PALs) and 183 6mer loci (91 PALs).

Targeting loci with higher numbers of repeats is often desirable because these more repetitive loci tend to show greater allelic variability (Kelkar *et al.* 2008). Of the more than 1800 PALs, 610 contained ≥ 6 repeats, 292 contained ≥ 8 repeats and 195 contained ≥ 10 repeats (Table 1). For all Nmer sizes except the more rare 6mers, there were tens to hundreds of PALs containing such longer tandem repeats (Table 1). Another relevant criterion for screening PALs that have higher probabilities of success is the uniqueness of the designed flanking primers, because this should predict the likelihood that a set of primers will exclusively amplify the locus of interest. The complete genome sequence of a *M. fulvius* is unavailable, and therefore, the absolute genomic copy number of designed primers is unknown. In its absence, PALFINDER uses the genomic shotgun sample set of genomic sequence to estimate the copy number of primers (in this approximately 7.7 Mbp data set). Because we sampled only a small fraction of the genome in this study (approximately 0.42%, based on a flow cytometry estimate of a approximately 1850 Mbp genome size in a related species, *M. lemniscatus*; MacCulloch *et al.* 1996), we do not expect that we identified all repetitive flanking sequences.

Table 1 Number of PALs that contain various numbers of tandemly repeated units, based on shotgun genomic sequencing via 454 of a single *Micrurus fulvius* animal

	Tandemly repeated units in PALs		
	≤ 10 repeats	≤ 8 repeats	≤ 6 repeats
2mers	67	100	222
3mers	39	59	104
4mers	54	85	199
5mers	31	42	76
6mers	4	6	9
Total	195	292	610

Previously (Castoe *et al*. 2012), we developed three progressively stringent criteria for screening primer copy number for such purposes (listed from least to most stringent): 1) the primer pair is observed only once in any single sequence read ('Pair'; Fig. 3), 2) one of the two flanking primers being observed only once in the entire data set ('Min (Fwd, Rev)'; Fig. 3) and 3) both primers being observed only once in the entire data set ('Fwd x Rev'; Fig. 3). Enforcing these criteria to filter PALs resulted in 1,554 PALs for the least stringent criterion (Pair) and 1090 for the most stringent criterion (Fwd × Rev). For the more frequent Nmers size classes, this filtering still yielded hundreds of candidate PALs (Fig. 3a).

Jointly using both flanking primer copy number and the number of tandemly repeated units as criteria for selecting likely variable and reliably amplifiable loci is a sensible approach. For example, for PALs with ≥ 6 tandem repeats, 1554 PALs meet the least stringent (Pair) cut-off, and 357 PALs meet the most stringent (Fwd × Rev; Fig. 3b). Thus, within any particular Nmer size class, under these joint selection criteria, there are still tens to hundreds of PALs to choose from (Fig. 3b). Furthermore, most studies do not have a need to restrict sampling to a single size class, whereas in the past, the requirement for selection by hybridization led to the use of a limited number of motifs that is no longer necessary given the Seq-to-SSR approach.

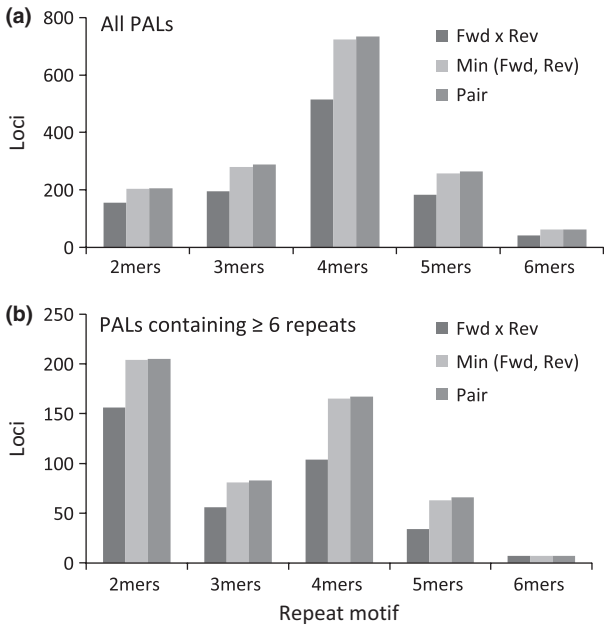

Fig. 3 The number of PALs meeting various criteria for flanking primer copy number or a combination of primer copy number and the minimum number of observed tandem repeats. (a) The number of PALs in which both primers are observed only once in the entire sequence data set (Fwd x Rev), at least one primer is observed only once in the entire sequence data set [Min (Fwd, Rev)], and the pair of primer sequences is observed together only once in any sequence read in the entire data set (Pair). (b) The number of PALs that contain 6 or more tandem repeats, and that also fit primer copy number criteria (above).

Comparing SSR landscapes between M. fulvius and other snakes

Prior to this study, similar 454 shotgun sampling was conducted for two other distantly related snake species – the Burmese python (*Python molurus*; Castoe *et al*. 2011, 2012) and the venomous copperhead (*Agkistrodon contortrix*; Castoe *et al*. 2009). Comparison of the frequencies of microsatellite loci of different Nmer classes per Mbp of genomic sequence shows substantial variation across the three snake species. All Nmer classes were more abundant in *Agkistrodon*, with intermediate levels in *Micrurus* and the lowest frequencies in the *Python* (Fig. 4). This trend is most pronounced for 2mer, 4mer and 5mer classes. *Agkistrodon* also has uniquely high 3mer frequencies compared to frequencies in *Micrurus* and *Python* that are similar to each other (Fig. 4).

To further compare the SSR landscapes of the three snake species, we visualized the number of microsatellite loci per Mbp for moderate to high abundance repeat sequence motifs (>5 loci per Mbp in any species; Fig. 5). Similar to comparisons across repeat classes (Fig. 4), most repeat motif sequences in *Micrurus* had numbers of SSR loci per Mbp intermediate between *Python* and *Agkistrodon*, with the *Agkistrodon* tending to have the highest frequencies of loci (Fig. 5). It has been shown that certain snake-specific transposable elements (TEs), Snake1 CR1 LINEs (chicken repeat 1, long interspersed

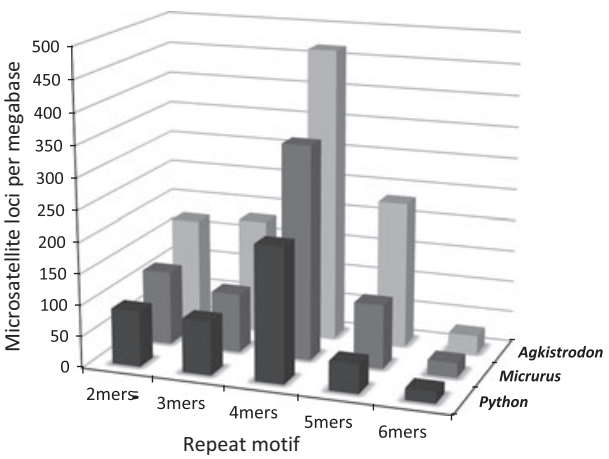

Fig. 4 Microsatellite locus length-class abundance per megabase in the genomes of three snake species sampled by 454 shotgun sequencing. *Micrurus* data are from this study, and *Agkistrodon* and *Python* data are based on previous studies (Castoe *et al*. 2009, 2012, respectively).

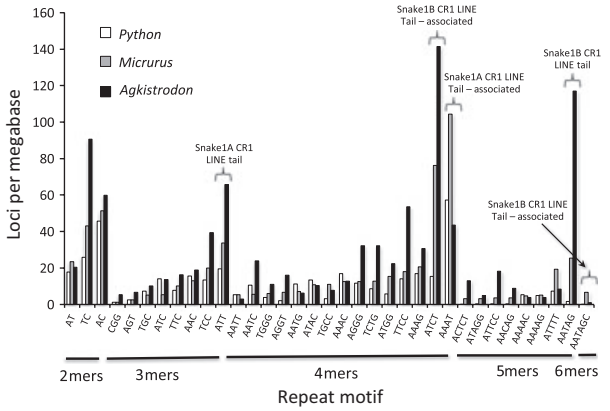

Fig. 5 Microsatellite locus abundance. The microsatellite abundance per megabase is shown for moderate-high abundance sequence motifs (>10 loci/Mbp) in the genomes of three snake species sampled by 454 shotgun sequencing. *Micrurus* data are from this study, and *Agkistrodon* and *Python* data are based on previous studies (Castoe *et al.* 2009, 2012, respectively). Sequence motifs that are known to be associated with the 3-prime tails of snake-specific transposable elements (Snake1A and Snake1B CR1 LINEs; Castoe *et al.* 2011) are indicated, as are high-frequency motifs that are one-step mutation variants of these elements (referred to as 'associated').

nuclear elements), are responsible for seeding both ATT (by Snake1A CR1) and AATAG (by Snake1B CR1) microsatellites in snake genomes; this is known to be extensive in the genome of *Agkistrodon*, but minimal in the *Python* genome (Castoe *et al.* 2011). Not surprisingly, these motifs are also the most abundant 3mer and 5mer motifs for both *Agkistrodon* and *Micrurus* (Fig. 5). Although 4mers do not appear to be directly seeded by TEs (i.e. they do not occur in the tails of active LINEs), 4mer sequence motifs that differ by one nucleotide from 3mer- and 5mer-seeded motifs are also high frequency in all three species, but especially in *Micrurus* and *Agkistrodon*. It is relevant to note that, because they are not differentiable, *PALFINDER* groups motifs with their reverse-complemented motifs and with motifs that are shifted by a base or more; thus, motif ATT is equivalent to TTA (shift) and TAA (reverse complement). Given this, it is notable that the highest frequency 4mer in *Micrurus* is AAAT (= ATTT), which is a one-mutation difference from ATT seeded by Snake1A CR1 LINEs (Fig. 5). Similarly, the highest frequency 6mer in *Micrurus* (AATAGC) and the second-highest frequency 4mer (ATCT = ATAG) are both one-mutation variants of the SSR motif (AAT AG) seeded by Snake1B CR1 LINEs (Fig. 5).

Efficient hierarchical screening of variable loci within the Micrurus fulvius complex

In the first round of PCR screening (for 40 tPALs in four individuals), we found that 20 loci were consistently amplifiable. In the second round of screening (76 individuals from across the geographic range of the *M. fulvius* complex), we were able to consistently amplify all 20 of these loci in 69 of the 76 individuals. In the outgroup species, *M. browni*, we were able to amplify several of these loci, but not consistently. We selected three 3mer and seven 2mer loci from among the 20 amplifiable loci for allelic size determination using fluorescent labelling. Eight of the ten loci scored for allelic length were polymorphic based on screening between 6 and 69 individuals (Table 2), including at least 50 individuals in five of these loci (L6, L7, L25, L35 and L40; Table 2).

We did not have any populations that were sampled for multiple individuals, which would have been ideal for validating loci by testing for Hardy–Weinberg equilibrium (HWE) of allelic frequencies. In lieu of this, we did test these five loci for HWE, based on grouping of individuals east and west of the Mississippi River. Of the five, three loci in the Western population (loci 25, 6 and 7) and two in the Eastern population (loci 40 and 7) significantly deviated ($P < 0.05$) from HWE; these deviations represented deficiencies in heterozygote genotypes. It is unclear, however, the degree to which the deviations from HWE genotypes might be due to population genetic structure or excessive inbreeding across the large ranges of the 'populations' we analysed, versus potential technical issues in scoring the data (null alleles, heterozygote genotype calling, etc.). This suggests that further evaluation of these loci with many samples from a single population would be preferable before these were used for extensive population genetic analyses to validate their performance.

To test the potential informativeness of the variation in the data set of five loci and 50 individuals (including 12 *M. fulvius*, 36 *M. tener*, 1 *M. bernadi* and 1 *M. tamaulipensis*), we used these data to perform a test of genetic isolation by geographic distance, and population structure analyses. Across this interspecific data set, there was a significant correlation between geographic and genetic distances ($y = 0.0019x + 3.4728$, $R^2 = 0.15$, $P = 0.010$; Fig. 6a). Log-likelihoods from the *STRUCTURE* analyses appeared to plateau after $K = 4$ populations (*lnL* scores were −230.8, −204.0, −174.3, −173.6, −176.2 and −176.9 for K values of 2–7, respectively). Interestingly, K values from 2 to 3 only weakly differentiated *M. fulvius* from *M. bernadi*, *M. tener* and *M. tamulipensis*. However, at $K = 4$ and beyond, *M. fulvius* alleles were distinctly clustered relative to the remaining members of the complex (Fig. 6b).

Discussion

We successfully implemented a rapid and efficient 3-step approach for identifying informative microsatellite loci

Table 2 Primer sequences and diversity for eight polymorphic microsatellite loci in the *Micrurus fulvius* complex

Locus (GenBank)	Primer sequence [Fluorophore label]	Repeat motif	N/N_a	Amplicon size range (bp)	H_O	H_E
L3	F: TTG TGT GCT GTT GCT GAT GG [HEX] R: AAT GCT CCC ACA TTT CTG GC	AC	33/6	267–281	0.24	0.32
L6	F: TTT ACA TTG GGA TGC TTG GC [HEX] R: TCC AAG TCC GAT TGG TTG G	AC	62/6	215–255	0.08	0.61
L7	F: TGT GCA CAC TTG AGA AAA GAT GC [6-FAM] R: GGA CAT CAG GCA AAT GCT GG	AC	68/6	159–169	0.31	0.53
L16	F: CAT CTC TCA GTG ACT CAT GTT GG [6-FAM] R:TCC AAA TAA AGC AGG TTT TCT CC	AC	6\2	256–260	0	0.28
L20	F: TGA CAC AGC AAT CAC AAA GAG C [HEX] R: TCT TAA CCA GTG GCT GTT GGG	AC	34/3	247–255	0.15	0.58
L25	F: TAT GCA AAA GTG GAG ACC CG [6-FAM] R: GCA GGG AGT GGG ATT AGA GC	TC	56/5	191–207	0.18	0.41
L27	F: AAA CTC TTC TTC TCC CAT AAC TCG C [6-FAM] R: CCT ACT TTA ATA GGC ACC TAA GAG G	TC	36/8	154–174	0.36	0.78
L40	F: ACC TGA ACC CAA AGG TCA CG [6-FAM] R: AGC CTT TCC TGC AAA TAC CG	TGC	69/6	164–187	0.07	0.15

F, forward primer; R, reverse primer; N, number of individuals scored; N_a, number of alleles; H_O, observed heterozygosity; H_E, expected heterozygosity.

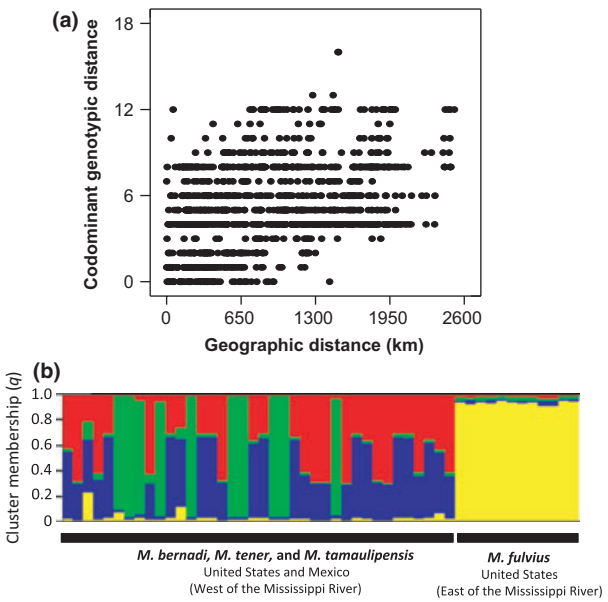

Fig. 6 Evidence that novel microsatellite loci are informative for population genetic questions across the *Micrurus fulvius* species complex. (a) Plot of Mantel test results for isolation by distance ($y = 0.0019x + 3.4728$, $R^2 = 0.15$, $P = 0.010$). (b) *STRUCTURE* analysis result with $K = 4$ populations, with posterior probability of population assignment for each individual shown along the Y-axis, and species designations per individual along the X-axis. The four classes were arbitrarily assigned the colours of yellow (high posterior for *M. fulvius* individuals) and green, red and blue (mixed posteriors for various *M. bernardi*, *M. tener* and *M. tamaulipensus*). Note that the green class tends to have a high posterior assignment, but does not correspond to a particular subspecies.

in a new nonmodel system. This approach utilized next-generation sequence reads and the Seq-to-SSR microsatellite identification approach (Castoe *et al.* 2009, 2012) to identify amplifiable microsatellite loci from unassembled raw shotgun genomic reads from a single individual. Employing a larger candidate set of inexpensive unlabelled primer pairs targeting select microsatellite loci, we then used a hierarchical strategy to screen for broad amplification, appropriate amplicon products and preliminary evidence of allelic size variation using agarose gels and fluorometric fragment analysis on the Agilent Bioanalyzer. Lastly, we scored allelic size variation among individuals in a subset of loci using fluorescently labelled primers on a capillary sequencer. This hierarchical approach was an economical and efficient means to obtain a useful set of species complex-specific informative microsatellite loci. Specifically, in terms of economy, the costs of this experiment included ca. $200 for library creation, $1350 for 1/8 of a 454-FLX sequencing run, plus bioanalyser chips for rapid sizing of microsatellite PCR amplicons (at $25/chip, accommodating 12 samples per) and the costs of PCR reagents and consumables.

The significant relationship between genetic and geographic distances (Fig. 6a) and the identification of population genetic structure related to a known dispersal barrier (Fig. 6b) together demonstrate that microsatellite loci identified from our library can be informative population genetic markers for *M. fulvius* and related species. The results from the *STRUCTURE* analyses are of particular interest regarding phylogenetic relationships within the *M. fulvius* complex. Whereas individuals can be iden-

tified as belonging to *M. fulvius* based on the *STRUCTURE* results, *M. bernadi*, *M. tener* and *M. tamaulipensis* individuals could not be definitively placed. If upheld by further studies (using additional taxa and loci), this result would call into question the validity of some currently recognized taxa.

In the past, microsatellite studies have concentrated mostly on shorter repeat motifs (e.g. 2–4mers) because of the need for hybridization-based approaches to preselect relatively frequent sequence motifs. Although 5mer and 6mer microsatellites are collectively nearly as common in the genome as 2–4mers, many individual 5–6mer sequence motifs are quite rare because there are so many more different possible sequence motifs for these longer Nmer repeats. Although particular 5–6mer motifs may be somewhat common, there is no way to definitively predict *a priori* for hybridization experiments which ones are more common than others. With the Seq-to-SSR approach, however, there is no need to perform laborious hybridizations, and therefore, there is no reason not to use 5mer and 6mer repeats. Shorter motifs may tend to be more variable, but practical issues related to scoring variability in shorter motifs can lead to uncertain motif lengths and thus to difficulty in interpreting results and possible misclassification of alleles. Future studies may benefit from utilizing longer repeat motif classes (e.g. 4mers and larger) to facilitate more accurate and rapid interpretation of allele lengths with automated methods, and inclusion of many loci in high-throughput approaches may obviate the need to focus on hypervariability.

Genomic SSR content is highly variable in snakes, and this is at least partially due to SSR expansion by transposable elements that vary in abundance across several snake genomes surveyed (Castoe *et al.* 2011; T. A. Castoe, A. P. J. de Koning & D. D. Pollock, unpublished data). Our results suggest that SSR loci in *M. fulvius* have intermediate genomic frequencies (approximately 1.4%) compared to *Python* (0.9%) and *Agkistrodon* (approximately 2.8%), particularly for 2–5mer class SSRs (Fig. 4). When SSR abundance across species is broken down by repeat motifs, this result generally still holds except for several extreme motifs (Fig. 5). Notable exceptions include the high-frequency Snake1A CR1 LINE tail-associated motif AAAT and the Snake1B CR1 LINE tail-associated motif AATAGC in *Micrurus* (Fig. 5). A particular type of CR1 LINE (Snake1) had apparently evolved SSRs at their 3-prime tails, which has led to the proliferation of certain SSR motifs in snake genomes that have experienced substantial CR1 LINE activity. Thus, certain motifs that occur on these tails (e.g. AAT, AATAGA), and related one-mutation-off variants of these, have become very frequent in the genomes of *Micrurus* and *Agkistrodon* compared to *Python* (Fig. 5). Although frequent, due to their association with transposable elements, the sequences flanking these repeats are often highly repetitive (observed often in our sequence sets) and therefore often are poor choices for targeted amplification and allele scoring (Castoe *et al.* 2011, 2012). Collectively, the difference in SSR motif landscapes among snake species augments previous evidence that such dynamics occurred in snakes (Castoe *et al.* 2011), and highlights the importance of species-specific approaches for marker identification (e.g. Seq-to-SSR).

As new sequencing technologies emerge and become mainstream, alternative approaches should be continually re-evaluated to maximize economy and efficiency. Our hierarchical approach to rapidly obtain variable microsatellite loci from a previously unstudied species appears to be a good example of how to leverage the strengths of existing technologies. We note that after this study began, an analogous Seq-to-SSR approach was developed that replaces 454 reads with Illumina paired-end reads. This reduces the cost of sequencing by more than 100-fold and is therefore capable of economically providing more information on repetitiveness of the flanking primers in the genome (Castoe *et al.* 2012). Increased sequencing enabled by using Illumina reads also provides a much larger set of PALs and thus allows enforcement of more stringent PAL selection criteria, such as primer copy number and the number of tandemly repeated units, to be used to select loci likely to produce reliable amplicons with allelic variation.

Acknowledgements

We acknowledge the support of the National Institutes of Health (NIH; R01-GM083127) to DDP, Bioclon to ENS, as well as a National Science Foundation collaborative grant to JAC (DEB-0102383), CLP (DEB-041600) and ENS (DEB-0416160). We thank S.B. Hedges for the specimen used for 454 sequencing. We thank individuals and institutions that contributed additional tissues for our study, including J. Vindum and R. Lawson (CAS), D. Dittmann (LSUMZ), M. Varela (Cuernavaca, Morelos), C. Franklin (Texas), O. Flores, A. Nieto and L. Canseco (MZFC), R. Murphy and R. McCulloch (ROM), D. Lazcano (UANL), D. Cannatella and G. Pauly (UT), K. Ray and P. Ustach (UTA), F. Mendoza-Paz (ITAH), T. Sinclair, T. Cole, T. LaDuc (UT), R. Aguilar (UAG), A. Carbajal-Saucedo (UAM), J. Dixon (TAMU), E. Sanchez (TAMUC), J. Reyes-Velasco (UTA), B. Stuart (NCMNS).

References

Burbrink FT, Castoe TA (2009) Molecular phylogeography of snakes. In: *Snakes: Ecology and Conservation* (eds Seigel R, Mullin S), pp. 38–77. Cornell University Press, Ithaca, New York.

Campbell JA, Lamar WW (2004) *The VEnomous reptiles of the Western Hemisphere*. Cornell University Press, Ithaca, NY.

Castoe TA, Poole AW, Gu W, de Koning APJ, Daza JM, Smith EN, Pollock DD (2009) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun sequence. *Molecular Ecology Resources*, **10**, 341–347.

Castoe TA, Hall KT, Guibotsy Mboulas ML *et al*. (2011) Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. *Genome Biology and Evolution*, **3**, 641–653.

Castoe TA, Poole AW, De Koning APJ *et al*. (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE*, **7**, e30953.

Del Giudice EM, Santoro N, Cirillo G, D'Urso L, Di Toro R, Perrone L (2001) Mutational screening of the CART gene in obese children: identifying a mutation (Leu34Phe) associated with reduced resting energy expenditure and cosegregating with obesity phenotype in a large family. *Diabetes*, **50**, 2157–2160.

Delmas CEL, Lhuillier E, Pornon A, Escaravage N (2011) Isolation and characterization of microsatellite loci in *Rhododendron ferrugineum* (Ericaceae) using pyrosequencing technology. *American Journal of Botany*, **98**, e120–e122.

Ersnt CH, Ersnt EM (2003) *Snake of the United States and Canada*. Smithsonian Press, Washington D.C.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines-recommendation for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources*, **6**, 1093–1101.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

Jennings TN, Knaus BJ, Mullins TD, Haig SM, Cronn RC (2011) Multiplexed microsatellite recovery using massively parallel sequencing. *Molecular Ecology Resources*, **11**, 1060–1067.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, **18**, 30–38.

Lavin-Murcio PA, Dixon JR (2004) A new species of coralsnake (Serpentes, Elapidae) from the Sierra de Tamaulipas, Mexico. *Phyllomedusa*, **3**, 3–7.

MacCulloch RD, Upton DE, Murphy RW (1996) Trends in nuclear DNA content among amphibians and reptiles. *Comparative Biochemistry and Physiology*, **113B**, 601–605.

Mikheyev AS, Vo T, Wee B, Singer MC, Parmesan C (2010) Rapid microsatellite isolation from a butterfly by *de novo* transcriptome sequencing: performance and a comparison with AFLP-derived distances. *PLoS ONE*, **5**, e11212.

Peakall R, Smouse PE (2006) GenAlEx 6: genetic analysis in excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

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

Rozen S, Skaletsky H (2000) Primer3 on the world wide web for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds Krawetz S, Misener S), pp. 365–386. Humana Press, Totowa, New Jersey.

Salazar AM, Vivas J, Sánchez EE *et al.* (2011) Hemostatic and toxinological diversities in venom of *Micrurus tener tener*, *Micrurus fulvius fulvius* and *Micrurus isozonus* coral snakes. *Toxicon*, **58**, 35–45.

Sánchez EE, Lopez-Johnston JC, Rodríguez-Acosta A, Pérez JC (2008) Neutralization of two North American coral snake venoms with Unites States and Mexican antivenoms. *Toxicon*, **51**, 297–303.

Santana QC, Coetzee MPA, Steenkamp ET, Mionyeni O, Hammond G, Wingfield M, Wingfield B (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217–223.

T.A.C. and J.W.S. designed experiments, conducted laboratory work, conducted analyses, and wrote the paper. E.N.S. and D.D.P. designed experiments and participated in writing the paper. M.J.I. and J.M.M. assisted with PCR and allele sizing. A.W.P. and A.P.J.D.K. contributed to software and analysis of the genomic data sample. C.L.P., E.N.S., and J.A.C. provided samples. All authors edited the manuscript.

Data Accessibility

The full set of microsatellite-containing raw 454 reads were deposited in GenBank under the accession numbers JX036543-JX040307. This microsatellite set is also available at www.snakegenomics.org/Raw_Data.html, as is the entire raw 454 shotgun sequence data set. Details of taxonomic and geographic sampling of specimens used for assessing microsatellite variation are provided in Online Table S1. Tabular output of microsatellite loci with flanking primers designed (PALs), with information on repeat sequence characteristics and empirically estimated primer copy number (in genomic sample) for primer sets for the *Micrurus fulvius* individual 454-sequenced, is provided in Online Table S2.

Supporting Information

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

Table S1 Taxonomic and geographic sampling of the *Micrurus fulvius* complex and related species.

Table S2 Summary information for microsatellite loci with flanking primers designed (PALs), with information on repeat sequence characteristics and empirically-estimated primer copy number (in genomic sample) for primer sets.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.