MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 341–347 doi: 10.1111/j.1755-0998.2009.02750.x

PERMANENT GENETIC RESOURCES ARTICLE

Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence

TODD A. CASTOE,* ALEXANDER W. POOLE,* WANJUN GU,*§ A. P. JASON DE KONING,* JUAN M. DAZA,*† ERIC N. SMITH‡ and DAVID D. POLLOCK*

*Consortium for Comparative Genomics, Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA, †Department of Biology, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816, USA, ‡Department of Biology & Amphibian and Reptile Diversity Research Center, The University of Texas at Arlington, Arlington, TX 76019, USA

Abstract

Optimal integration of next-generation sequencing into mainstream research requires re-evaluation of how problems can be reasonably overcome and what questions can be asked. One potential application is the rapid acquisition of genomic information to identify microsatellite loci for evolutionary, population genetic and chromosome linkage mapping research on non-model and not previously sequenced organisms. Here, we report on results using high-throughput sequencing to obtain a large number of microsatellite loci from the venomous snake *Agkistrodon contortrix*, the copperhead. We used the 454 Genome Sequencer FLX next-generation sequencing platform to sample randomly ~27 Mbp (128 773 reads) of the copperhead genome, thus sampling about 2% of the genome of this species. We identified microsatellite loci in 11.3% of all reads obtained, with 14 612 microsatellite loci identified in total, 4564 of which had flanking sequences suitable for polymerase chain reaction primer design. The random sequencing-based approach to identify microsatellites was rapid, cost-effective and identified thousands of useful microsatellite loci in a previously unstudied species.

Keywords: high-throughput marker identification, next-generation sequencing, simple sequence repeats, snake genomics, Viperidae

Received 14 April 2009; revision received 4 June 2009; accepted 12 June 2009

Introduction

The accessibility of next-generation high-throughput sequencing technology has transformed life sciences research, providing a massive cost reduction per nucleotide sequenced (see Hudson 2008, for review). In many cases, this revolution calls for the redesign of previous experimental methods that employ laborious techniques. The extremely high mutation rates and simple codominance of microsatellite loci make them the marker of choice for medium to fine scale population genetic studies, although the costs of developing useful markers can be prohibitive. For example, traditional approaches for the identification of microsatellite loci from non-model species require substantial technical effort to create enriched microsatellite libraries, including cloning, hybridization to detect positive clones, plasmid isolation and Sanger sequencing. As microsatellites often do not transfer well from one species to another, this process may need to be repeated *de novo* for each species studied, thus imposing a substantial burden for initiating research on previously unstudied species. Given the ever-lower costs of next-generation sequencing approaches, it seems

Correspondence: David D. Pollock, Fax: +1 303 724 3215;
E-mail: david.pollock@ucdenver.edu
§Present address: Key Laboratory of Child Development and Learning Science, Southeast University, Ministry of Education, Nanjing 210096, China.
The first two authors contributed equally.

feasible that the overall cost of obtaining microsatellite loci sequences may be made cheaper and more efficient by incorporating next-generation sequencing. We suggest that microsatellite detection can be easily achieved through (random) shotgun genome sequencing and candidate microsatellite loci can be identified from the set of randomly sampled shotgun reads. This potential has also been recently recognized and explored by others (Abdelkrim *et al.* 2009; Allentoft *et al.* 2009).

Standard enrichment-based approaches for isolating microsatellite loci require *a priori* choices about what types of microsatellite loci to target (both repeat size, and repeat motif sequence) because specific oligonucleotide probes are used. This inherently biases which subset of microsatellite loci are identified and ultimately limits the diversity of different microsatellite types that are used in genotyping to a small subset of all possible repeat sizes and motifs. Furthermore, given the great variation in the abundance of different sizes and sequence motifs of microsatellite repeats in different species (Tóth *et al.* 2000; Abdelkrim *et al.* 2009; Santana *et al.* 2009), uninformed *a priori* choices about which motifs to target may lead to limited success in obtaining sufficient numbers of useful microsatellite loci. In contrast, microsatellite identification from randomly sequenced genomic regions (i.e. shotgun library sequencing) allows for unbiased surveys of all types of microsatellite loci present in a genome (in frequencies presumably proportional to their genomic abundance); a diversity of repeat motif types can be identified and used to survey genotypic diversity. In addition, this approach amasses novel genomic resources that may be useful for both related and unrelated research.

Here, we report on results obtained by applying this approach to detect microsatellite loci in the venomous snake *Agkistrodon contortrix* (the copperhead). This species ranges throughout much of the eastern half of the United States (Campbell & Lamar 2004), and has an estimated haploid genome size of 1.33 Gbp (De Smet 1981). In this study, we used the 454 Genome Sequencer FLX platform (454 hereafter) to sequence ~27 Mbp of sheared genomic DNA from *A. contortrix*, thus randomly sampling about 2% of the haploid genome of this diploid species. From this random sample, we identified thousands of microsatellite loci and the subset of these loci that contained flanking sites suitable for polymerase chain reaction (PCR)-based amplification and scoring for population genetic studies.

Methods

Preparation of shotgun libraries for 454 sequencing

A single specimen of the copperhead, *Agkistrodon contortrix*, from Val Verde Co., Texas, USA, was used as the sole source of tissue/DNA for this study; the voucher specimen was deposited at The University of Texas at Arlington Amphibian and Reptile Diversity Research Center (UTA R-55729). DNA was extracted using a DNEasy kit (Invitrogen) from liver tissue that had been frozen at −80 °C after being snap-frozen in liquid nitrogen, using the optional RNase treatment option of the protocol. Multiple DNA isolations were pooled, precipitated with ethanol/sodium acetate, washed once with 70% ethanol, and resuspended in TE buffer. A total of 5 μg of this DNA was used in the 454 FLX shotgun library preparation following the manufacturer's protocols and quality control steps. This shotgun library was ultimately sequenced using the 454 FLX GS LR70 (100 cycle) sequencing kit and protocol on a 70 × 75 cm picotiterplate. This library was used in several nonoptimized diagnostic trial experiments, sequenced on both ¼ and ⅛ 70 × 75 cm picotiterplate regions, yielding ~27 Mbp in total. The same amount of data can often be obtained from a single ¼ plate optimized run and newer 454 'titanium' runs should yield more and longer sequence in a ⅛ plate run (see further discussion below).

Identification of microsatellite loci

A Perl script (Appendix S1) was written to extract reads that contained perfect dinucleotide, trinucleotide, and tetranucleotide tandem repeats (i.e. microsatellite loci) that were at least 12 bp in length (e.g. six tandem repeats for dinucleotides). The reads were then sorted by the monomer sequence of the repeat (e.g. TAC or TA repeats) and by the number of tandemly repeated units. Reverse-complement repeat motifs (e.g. TG and CA) and translated or shifted motifs (e.g. TGG and GTG) were grouped together such that there were a total of four unique dinucleotide repeats and 10 and 33 unique three and four-nucleotide repeats respectively. If multiple microsatellite loci were discovered in a single read, the locus was considered a compound repeat if the microsatellites had different motifs and was considered a broken repeat if the microsatellites had the same motif. For most analyses, the locus with the greatest number of perfectly repeated units was used to classify that read and the other microsatellites in that read were not counted.

Screening of loci suitable for PCR and automated design of amplification primers

Newly identified microsatellite loci are typically useful only if primers in the nonrepeated flanking regions around the microsatellite can be designed and used successfully for PCR amplification. We therefore screened reads with microsatellite loci for flanking regions with high quality PCR priming sites; we refer to such loci as

'potentially amplifiable loci' or PAL. The primer-pair design process was automated to submit large batches of sequences to a local installation of the program Primer 3 (Rozen & Skaletsky 2000) using a Perl script (Appendix S1). Low complexity and simple repeat sequences were masked using the Repbase v14.01 database (Jurka *et al.* 2005) so that they would not be allowed as primer sites. We used fairly stringent criteria for primer design, including the following specifications: (i) GC content >30%; (ii) melting temperatures 58–65 °C with a maximum 2 °C difference between paired primers; (iii) the last two 3′ nucleotides were G or C (a GC 'clamp'); (iv) minimum amplicon length of 60 bp; and (v) maximum poly-N of four nucleotides. All other parameters were set to Primer 3 default values. If all criteria were met, a single primer-pair was chosen based on the highest Primer 3 assigned score (Rozen & Skaletsky 2000) and targeting the longest microsatellite element within a sequence. The Perl script used to identify and classify microsatellite reads and automate primer design is provided as Appendix S1.

Vertebrate genomes comprise a great deal of repetitive DNA that may confound PCR amplification and scoring of identified microsatellites. To assess the possible repetitive origins of some of our primers, we counted the number of times the designed amplification primers were found in all reads in the entire sequence data set and the number of reads in which a pair of amplification primer sequences was in the correct orientation to produce a PCR amplification product. This is useful to help predict the probability of successful amplification of an amplicon of interest; multiple priming sites can produce multiple amplification products or reduce the quality of amplifications due to primer non-specificity. For paired microsatellite amplification primers repeated among reads, we separately evaluated whether the amplicon and the entire read were identical. To calculate the expectation that two reads overlapped, we conservatively assumed that they would have to overlap by at least 100 bp to lead to identification of the same microsatellite with viable flanking PCR primer sites. Assuming a genome size of 1.33 Gbp and average read lengths of 212 bp, the probability that any two reads overlap each other for 100 bp of the same genomic sequence is 112/1 330 000 000, or 8.42×10^{-8}. Cumulatively, the expectation for the number of PAL sequenced more than once is less than one (0.88).

Results

A total of 128 773 reads were obtained from the 454 shotgun library of *Agkistrodon contortrix*. The mean read length was 215 bp and a total of 26 874 213 bp of sequence was obtained. The average GC content of these data was 42.51%. We identified 14 612 reads (11.3% of all reads) that contained our definition of microsatellite loci (Table 1). There were 3323 reads that contained dinucleotide repeats, 3606 that contained trinucleotide repeats and 7683 that contained tetranucleotide repeats (Fig. 1). These reads containing identified microsatellites were deposited in GenBank under Accession numbers GQ184828–GQ199439.

To estimate the number of loci that represent promising candidates for PCR amplification-based scoring of microsatellite length variation, we screened loci to determine which contained suitable flanking PCR primer sites. We identified 4564 such 'potentially amplifiable loci' (PAL). Reads containing PAL represented 3.5% of all reads, and 31.2% of reads identified as microsatellites (Table 1). Comparing across the three classes of repeats, tetranucleotide repeats had the greatest numbers of loci and the most PAL (7683 and 2376 respectively) and dinucleotides the least (3323 and 701 respectively). Within repeat classes, trinucleotides had the greatest percentage of PAL (41.2%), and dinucleotide repeat loci had the lowest percentage (21.1%; see also Fig. 1).

Table 1 The numbers of microsatellite loci identified, and the subset of these that are potentially amplifiable (containing suitable PCR priming sites) in 128 773 reads from 454 sequencing

	Number of repeats per loci	Number of loci identified	Number of potentially amplifiable loci	Percent of loci potentially amplifiable
Dinucleotides	All (≥6 repeats)	3323	701	21.1
	>10 Repeats	1178	144	12.2
	>20 Repeats	270	12	4.4
Trinucleotides	All (≥4 repeats)	3606	2119	41.2
	>10 Repeats	736	206	27.9
	>20 Repeats	170	66	38.8
Tetranucleotides	All (≥3 repeats)	7683	2376	30.1
	>10 Repeats	1858	80	4.3
	>20 Repeats	144	9	6.3
Total		14 618	4564	31.2

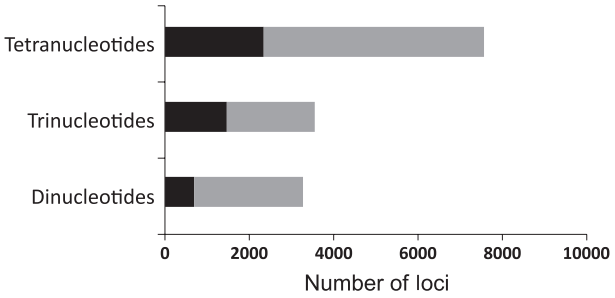

Fig. 1 Numbers of identified microsatellite repeat loci (grey), and the number of these loci with suitable flanking PCR-primer sites (potentially amplifiable loci, PAL; black), in 128 773 reads randomly sampled from the *Agkistrodon contortrix* genome by 454 sequencing.

There are large differences in the relative abundances of specific repeat motifs (Fig. 2). TC was the most frequent dinucleotide repeat and the most frequent repeat overall, while GC was the least frequent dinucleotide repeat (Fig. 2A). AC repeats had the most amplifiable loci (Fig. 2A). Trinucleotide repeat motifs were dominated by AAT repeats, with TCG being the least frequent (Fig. 2B). Unlike the case for dinucleotide repeats, observed counts of repeat-containing reads within a trinucleotide motif class were largely proportional to the number of PAL per motif (Fig. 2B). For tetranucleotides, ATCT, AAAT and TTCC were the three most common motifs (Fig. 2C); note that half of all tetranucleotide motifs were observed fewer than 30 times (Fig. 2C). As with dinucleotide repeats, the most frequent tetranucleotide motif (ATCT) contained relatively few PAL and there was fairly low correspondence between the number of loci observed and the number of PAL per motif (Fig. 2C).

The relative numbers of perfect tandemly repeated units for each microsatellite locus length class were fairly similar across the three classes of repeats (Fig. 3). In all three classes of repeats, there were many short repeats (those with the fewest repeat unit counts), and fewer long repeats (Fig. 3). Also, smaller proportions of the longer loci were potentially amplifiable (Fig. 3; Table 1). This is expected because as the size of the microsatellite region increases, the amount of flanking sequence decreases; thus, there is less potential flanking sequence available for primer design with longer microsatellite loci. This inverse relationship between the number of perfectly repeated units and the percent of reads that are potentially amplifiable is also clear from Table 1. For each class of microsatellite, the proportion of PAL with repeats of more than 20 units drops sharply (particularly for dinucleotide and tetranucleotide repeats; Table 1). Many of the potentially amplifiable regions (between designed primer pairs) contain longer stretches of imperfect (broken) or compound repeats, however, and the total number of tandem repeats in the potentially amplifiable regions has a much longer tail (Fig. 4). Of the 14 612 total microsatellite reads identified, 3201 contained compound repeats and 1444 contained broken repeats.

Another question, given the random sequence sampling approach used, is whether any locus was sequenced twice. If PAL are unique in the genome, the probability that any of the 4564 PAL were sampled multiple times is quite small (<0.88; see Methods). There are, however, 62 primer pairs with duplicate copies in the

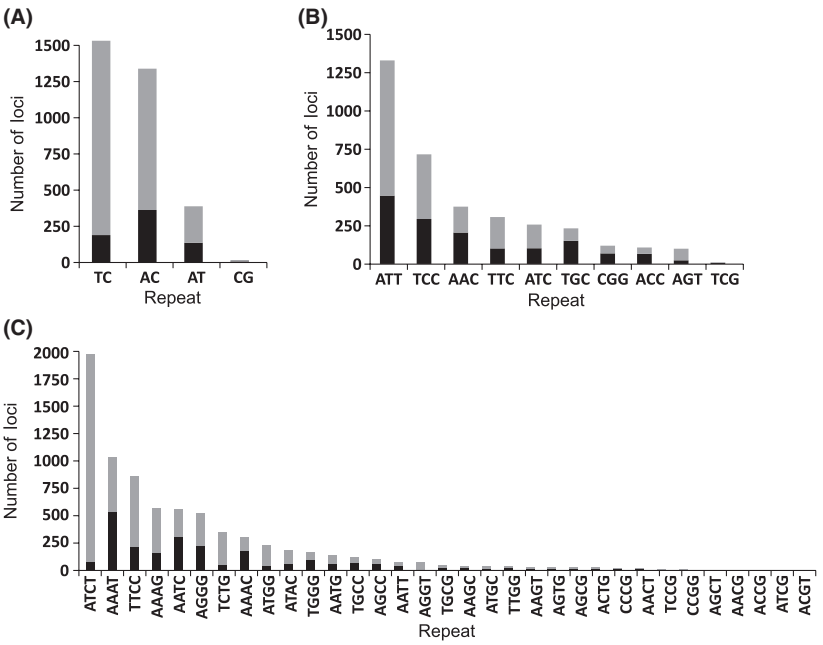

Fig. 2 Observed counts of identified microsatellite loci (grey), and the subset of these that contain PCR-primer sites (potentially amplifiable loci, PAL; black) for different repeat sequence motifs of (A) dinucleotide repeats, (B) trinucleotide repeats and (C) tetranucleotide repeats.

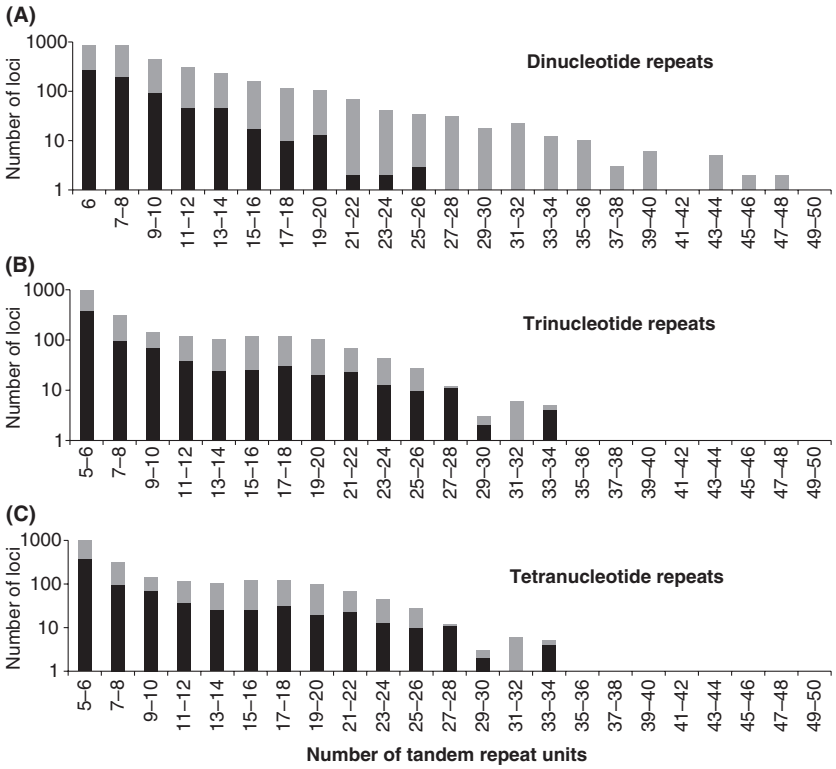

Fig. 3 Counts of the number of perfect tandem repeat units per identified (grey), and per potentially amplifiable (black) microsatellite locus, in 128 773 reads randomly sampled from the *Agkistrodon contortrix* genome by 454 sequencing. Counts of perfect tandem repeat units are given for (A) dinucleotide repeats, (B) trinucleotide repeats and (C) tetranucleotide repeats. Note that the *Y*-axis in all three graphs is on a logarithmic scale.

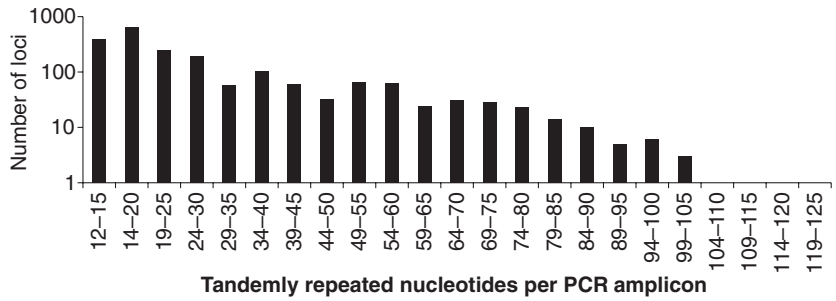

Fig. 4 The total number of tandemly repeated nucleotides within the bounds of designed PCR primers, including all broken and compound repeats as well as perfect repeats.

PAL data set (Appendix S2 and Table S1). Interestingly, 15 of these 62 primer pairs had identical or nearly identical reads associated. These identical reads may be from the same locus, or from multiple copies from an undiverged transposable element or tandem duplication or may indicate a technical replication of unknown cause in the 454 sequencing process. We note that approximately nine of the 14 618 identified microsatellite loci are expected to have been sequenced more than once with an overlap of 100 bp or more, but this number is too small to have a substantial impact on our analysis of the frequencies of different microsatellite types.

Of much greater importance is the likelihood that some of the primers designed are derived from repeated sequences such as transposable elements. To predict this (without knowing the repeat structure of the *A. contortrix* genome), we counted the number of exact matches to each primer sequence in each read in the entire data set (not just the amplifiable loci) and also counted the number of reads in which an amplification primer pair was found in the correct orientation to produce a PCR product. Two-thousand eight-hundred and seven of the 4564 PAL have only one copy of each primer in the entire sequence data set and are thus most likely to be of unique origin and to amplify a single locus. A further 1252 have more than one copy of only one of the two primers in the entire sequence data set. Perhaps the least likely to amplify a single locus are the are the 505 PAL with more than one copy of both primers, particularly the 246 PAL whose primer pair occurs together in other reads in the correct orientation (for producing a PCR amplicon;

see Table S1). These 246 almost certainly represent microsatellies within transposable element regions.

In Appendix S2, we have assembled a summary of the results containing the unique sequence accession of each read, the repeat monomer sequence, the number of perfect tandemly repeated units and the forward and reverse primer sequences (each with unique names) for each PAL. This file also includes counts of the frequency that each amplification primer was found in the entire shotgun read database, the number of times that a primer pair was found in the correct orientation for producing an amplification product, and indicated PALs that appear to be the same locus (same primer set, same amplicon sequence). We have also included a text file of all microsatellite reads (Appendix S2) indexed by their unique sequence accession to facilitate rapid reference between reads and Table S1.

Discussion

Using fairly small amounts of next generation DNA sequencing, we identified an extensive set of 14 612 microsatellite loci (4564 of which are potentially amplifiable via PCR) without previously knowing anything about the target genome sequence. This unbiased set of PAL and their primers is available for evolutionary and population genetic research (e.g. Goldstein & Pollock 1997) as well as high-resolution chromosome linkage mapping studies. Compared to the weeks or even months that can be spent obtaining only tens of microsatellite loci by traditional approaches, the thousands identified here required only ~9 days (some partial) to take the sample from tissue through DNA extraction, library creation, library titration and sequencing. Since these data were collected, there has been another drop in sequencing costs with the release of the 454 FLX-XLR Titanium chemistry. These new reagents provide an expected fivefold increase in sequence output and twofold longer reads than the LR70 sequencing kit we used in this study. Thus, the present experiment could be repeated (with additional data collected) today for about $2000 in reagent costs, using a ⅛ region of a Titanium 70 × 75 picotiterplate and Titanium sequencing chemistry. The longer reads of the Titanium sequencing reagents should also yield higher percentages of PAL and be able to identify longer microsatellites that are likely to be highly variable (see also Allentoft *et al.* 2009). As moderately long reads were central to our strategy for identifying candidate microsatellite loci, it is difficult to envision how short-read next-generation sequencing approaches (e.g. Illumina) could be readily used for an analogous approach.

Our results suggest that this random genomic sequencing approach represents a rapid and cost effective means to identify thousands of potential amplifiable and variable microsatellite loci in a previously unstudied species. In addition to identifying microsatellite loci, we filtered the data to find the subset of microsatellite-containing reads for which PCR primers could be designed and could thus be readily screened for variation with further PCR. Overall, about one third of the microsatellite loci had satisfactory primer design sites, yielding >4000 microsatellite loci that can potentially be used to amplify and score microsatellite alleles based on length variation (repeat number). Although this approach did not provide detailed information on which microsatellite loci are variable, and thus suitable for population genetic research, the sheer number of loci identified allow the preferential targeting of loci that would be expected to be highly variable. For example, perfect (vs. imperfect or compound) microsatellite repeats (Buschiazzo & Gemmell 2006), as well as longer microsatellite repeats (Kelkar *et al.* 2008) are generally known to exhibit greater allelic variability. Thus, screening of the longer perfect repeat loci would be expected to yield a high proportion of loci that are variable within and between populations. We also used the large shotgun read data set to identify which amplification primers are unique to a PAL, vs. repeated in many other sequence reads (possibly due to its association with a transposable element). Using this information, PAL that have more unique primer sequences and did not have primer pairs occurring in other reads would be a priority for further screening because of their higher probability of producing successful locus-specific PCR amplification products. Only 246 of these matched primer pairs are found more than once in the entire data set, indicating that these are probably internal to a repeat element, and are thus poor choices for amplification. This identification of microsatellite loci with regions that may be associated with repeat elements is an excellent example of a useful application of the nonmicrosatellite by-catch of this random sequencing approach.

Compared to traditional methods of microsatellite identification, the random shotgun approach rapidly and inexpensively produced a large number of diverse candidate microsatellite loci. This method of essentially screening the entire genome for candidate microsatellite loci may provide a viable and preferable alternative to targeted enrichment-based approaches for obtaining candidate loci (Väli *et al.* 2008). Recently, others have taken an alternative approach by applying standard microsatellite enrichment techniques to prepared 454 shotgun libraries, and 454 sequencing the resulting enriched library (Santana *et al.* 2009). As discussed above, there are disadvantages to the enrichment process (i.e. biased microsatellite discovery) but the obvious advantage is that less sequence is wasted on nonmicrosatellite sequences (Santana *et al.* 2009). If new methods became available to create barcoded shotgun libraries for many individuals

more rapidly and cheaply, it may soon be possible to conduct experiments that simultaneously identified and screened hundreds of individuals for hundreds of loci simultaneously in a single 454 run, for example. In general, these types of approaches for identifying (and ultimately scoring) large numbers of microsatellite loci would tremendously increase the power and resolution of population genetic studies (Selkoe & Toonen 2006).

The challenge of harnessing the power and ideal price-point of next-generation sequencing is a balance of two factors: the first is increasing economy by developing methods for targeting more precisely the sequences of interest and the second is to identify other bottlenecks of time or monetary expense elsewhere in research programmes and use this sequencing power to replace or circumvent these with alternative approaches (as in the current method). This and other recent adaptations of existing methods in molecular ecology to incorporate next-generation sequencing into microsatellite identification (Abdelkrim *et al.* 2009; Allentoft *et al.* 2009; Santana *et al.* 2009), high-throughput diet analysis (Valentini *et al.* 2009), transcriptome characterization of non-model species (Vera *et al.* 2008) and high-throughput surveying of microbial ecosystem diversity (Edwards *et al.* 2006; Hamady *et al.* 2008) represent new and exciting transformations in the depth and scope of questions that can be addressed with such economical access to tremendous sequencing capabilities (see also Ellegren 2008).

Acknowledgements

We acknowledge the support of the National Institutes of Health (NIH; GM065612-01, GM065580-01) to DDP, and an NIH training grant (LM009451) to TAC. We thank C. Franklin (UTA) for the specimen used for this study.

References

Abdelkrim J, Robertson BC, Stanton JL, Gemmel NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, **46**, 185–192.

Allentoft ME, Schuster SC, Holdaway RN *et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *BioTechniques*, **46**, 195–200.

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays*, **28**, 1040–1050.

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

De Smet WHO (1981) The nuclear Feulgen-DNA content of the vertebrates (especially reptiles), as measured by fluorescence cytophotometry, with notes on the cell and chromosome size. *Acta Zoologica et Pathologica Antverpiensia*, **76**, 119–167.

Edwards RA, Rodriguez-Brito B, Wegley L *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**, 57.

Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17**, 1629–1631.

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods in phylogenetic inference. *Journal of Heredity*, **88**, 335–342.

Hamady M, Walker JJ, Harris JK, Gold N, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, **5**, 235–237.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Jurka J, Kapitonov VV, Pavlicek A *et al.* (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research*, **110**, 462–467.

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

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, NJ.

Santana QC, Coetzee MPA, Steenkamp ET *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217–223.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Tóth G, Gáspári GZ, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, **10**, 967–981.

Valentini A, Miguel C, Nawaz MA *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51–60.

Väli Ü, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808–3817.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Supporting Information

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

Appendix S1. Perl script used to identify microsatellite loci and suitable flanking PCR primer sites in 454 reads.

Appendix S2. Sequence reads that contained identified microsatellite loci.

Table S1. Details of microsatellite locus characteristics and designed primers for each microsatellite locus identified.

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