Anonymous nuclear loci in non-model organisms: making the most of high-throughput genome surveys

Terry Bertozzi1,2,*, Kate L. Sanders3, Mark J. Sistrom4 and Michael G. Gardner5,6
1Evolutionary Biology Unit, South Australian Museum, North Terrace, Adelaide, SA 5000, Australia, 2School of Molecular and Biomedical Science, North Terrace, Adelaide, SA 5000, Australia, 3Ecology, Evolution and Landscape Science, University of Adelaide, North Terrace, Adelaide, SA 5000, Australia, 4Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520-8106, USA, 5School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia and 6Australian Centre for Evolutionary Biology and Biodiversity, School of Earth and Environmental Science, University of Adelaide, Adelaide, SA 5000, Australia

ABSTRACT
Motivation: When working with non-model organisms, few if any species-specific markers are available for phylogenetic, phylogeographic and population studies. Therefore, researchers often try to adapt markers developed in distantly related taxa, resulting in poor amplification and ascertainment bias in their target taxa. Markers can be developed de novo and anonymous nuclear loci (ANL) are proving to be a boon for researchers seeking large numbers of fast-evolving, independent loci. However, the development of ANL can be laboratory intensive and expensive. A workflow is described to identify suitable low-copy anonymous loci from high-throughput shotgun sequences, dramatically reducing the cost and time required to develop these markers and produce robust multifocus datasets.

Results: By successively removing repetitive and evolutionary conserved sequences from low coverage shotgun libraries, we were able to isolate thousands of potential ANL. Empirical testing of loci developed from two reptile taxa confirmed that our methodology yields markers with comparable amplification rates and nucleotide diversities to ANLs developed using other methodologies. Our approach capitalizes on next-generation sequencing technologies to enable the development of phylogenetic, phylogeographic and population markers for taxa lacking suitable genomic resources.

Contact: terry.bertozzi@samuseum.sa.gov.au

Received on November 21, 2011; revised on April 20, 2012; accepted on May 5, 2012

1 INTRODUCTION
In a recent review, Thomson et al. (2010) noted that next-generation sequencing methodologies make the acquisition of many-marker datasets readily feasible. The ability to generate libraries of random sequence from across the genome at a relatively low expense has already changed the way that markers are developed for phylogenetic and population genetic studies. In particular, the use of the Roche 454 platform for generating large fragments of randomly sheared genomic DNA has already provided a wealth of markers. Most marker development studies have focussed on the large number of microsatellites that can be readily obtained from partial sequencing runs on non-model organisms (e.g. Abdelkrim et al., 2009; Allentoft et al., 2009), and large numbers of species can be analysed simultaneously on this platform (Gardner et al., 2011). In addition, almost entire mitochondrial genomes and a number of protein-coding genes (Rasmussen and Noor, 2009) have been recovered from these kind of data. One class of marker that should be present in high numbers but has yet to be exploited from these data are anonymous nuclear loci (ANL) (Karl and Avise, 1993).

ANL represent randomly selected fragments of DNA that, by chance, are likely to be from non-coding regions of the genome, given that a large percentage of eukaryotic genomes comprise non-coding regions (Thomson et al., 2010). The advantages of these markers are numerous: they are likely to be independent; not biased to a single area of the genome and unlikely to be under selection and so free to accumulate mutations at varying rates, providing a suite of sufficiently fast-evolving markers useful for phylogenetics, phylogeography and population genetics. ANL have been successfully used in multilocus reconstruction of phylogenetic (Thomson et al., 2008) and phylogeographic (Rosenblum et al., 2007; Lee and Edwards, 2008) histories. A criticism of using anonymous markers is that paralogous loci and repetitive elements may go undetected, potentially confounding phylogenetic and phylogeographic inferences (Thomson et al., 2010). In addition, the development of these markers has required intensive laboratory methods including a substantial amount of cloning during development from small insert libraries (Lowe et al., 1998; Jennings and Edwards, 2005), Bacterial Artificial Chromosome libraries (Thomson et al., 2008) and Amplified Fragment Length Polymorphism fragments (Brugmans et al., 2003). Herein, we describe the development of ANL from randomly generated sequences produced by the Roche 454 platform and provide empirical data to evaluate the methods.

2 METHODS
2.1 Marker development
Sequence reads were generated for two squamate reptiles, the gecko Gehyra lazelli and the sea snake Hydrophis spiralis, by sequencing randomly sheared genomic DNA on the Roche 454 GS-FLX titanium platform at the Australian Genomic Research Facility, Brisbane, Australia, following the methodology of Gardner et al. (2011). Approximately, one-eighth of a titanium flow cell for each species returned a total of 87 899 reads for G. lazelli (average length...
A workflow was designed to successively filter out sequence reads of non-interest (i.e., repetitive, highly conserved and coding sequences) using software available in the public domain. All analyses were run on an Intel quad core 2.4 Ghz PC with 8 GB RAM running Fedora Linux (Release 11). When possible, parallel processing capabilities using multiple processor cores were used to speed up analyses. Default parameter values were utilized for all software unless otherwise stated. The sequences for each species were analysed separately. FASTA formatted sequence reads were extracted from the GS-FLX output files using the ‘sffinfo’ program, available as part of the Roche GSassembler suite. RepeatScout (Price et al., 2005) was used to identify highly repetitive sequences in the data after running the build, limer_table script, which is included with the package. The repeat library generated contained high copy number sequences including mitochondrial fragments, simple sequence repeats (microsatellites), transposable elements and species-specific repeats. This repeat library was used as a ‘custom library’ in RepeatMasker (Smit et al., 1996–2010) to mask these highly repetitive sequences in the original FASTA file. In addition, repeats included in the Repbase Update (Jurka et al., 2005) ‘repeatmaskerbibliotities-20090004’ were also used with RepeatMasker to remove known repeats that were present in low copy number. The parallel processing option was enabled and the bacterial insertion element check skipped (using the ‘–no_is’ option). Masked areas were removed from sequences using the SeqClean PERL script (available from http://compbio.dfci.harvard.edu/tgi/software) enabling the parallel processing option. The SeqClean script was run three times with the minimum ‘clear range’ (-l option) set to 250, 350 and 450 bp, respectively, parallel processing option. The SeqClean script was run three times with the minimum ‘clear range’ (-l option) set to 250, 350 and 450 bp, respectively, to determine the number of sequences of each size class available for further analysis.

To remove highly conserved sequence or known coding regions, the resulting reads were compared to existing GenBank sequences using BLASTN (version 2.2.21). The makeblastdb package included in the BLAST distribution was used to create local, searchable BLAST databases comprising available messenger RNA (mRNA) and EST sequences from the family Elapidae, which includes the subfamily Hydrophiinae, to compare to the H. spiralis reads and similarly mRNA and EST sequences from the genus Gehyra to compare to the G. lazelli reads. An additional BLAST database constructed from the square-nosed Aonuilus cariniformis genome (AnoCar 1.0, February 2007, Broad Institute) was also used to screen each dataset. Finally, matches to mitochondrial sequences from Rangouas fasciatus (Elapidae: GenBank accession NC_011393) and Gehyra gekko (Gekkonidae: GenBank accession NC_007627) were determined to remove any low copy number mitochondrial sequences not removed by the RepeatScout program. High similarity matches, defined as e-value < 10^-10, were output in tabular format and removed from the FASTA files using an in-house PERL script. A PERL script automating the analysis pipeline to this point is available from http://www.samuelsum.sa.gov.au/assets/files/science/ evolutionary-biology-unit/anonmarker.tar.gz.

Primer design was carried out using the program ‘Primer3’ (Rozen and Skaletsky, 2000), as implemented in software package Geneious, by specifying a product size between 400 and 500 bp and using default setting for optimal primer size (20 bp), Tm (60°C), %GC content (50%) and minimizing hairpins and primer dimers.

### Table 1. Number of sequences with lengths > 250, 350 and 450 bp after the removal of repetitive and high copy number sequences

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Contiguous sequence length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥250</td>
</tr>
<tr>
<td></td>
<td>≥350</td>
</tr>
<tr>
<td></td>
<td>≥450</td>
</tr>
<tr>
<td>Gehyra lazelli</td>
<td>29 054</td>
</tr>
<tr>
<td>Hydrophis spiralis</td>
<td>32 085</td>
</tr>
</tbody>
</table>

### Table 2. Number of sequences ≥450 bp remaining after further filtering and automated PCR primer design

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sequences ≥450 bp</th>
<th>Sequences ≥450 bp after filtering</th>
<th>PCR product &gt;400 bp</th>
<th>PCR product &gt;500 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gehyra lazelli</td>
<td>7 957</td>
<td>5 560</td>
<td>3 874</td>
<td>2 45</td>
</tr>
<tr>
<td>Hydrophis spiralis</td>
<td>7 733</td>
<td>5 822</td>
<td>4 398</td>
<td>7 6</td>
</tr>
</tbody>
</table>

### 3 RESULTS

#### 3.1 Marker development

The number of sequences for each taxon with contiguous sequence lengths > 250, 350 and 450 bp after masking repeats and other highly repetitive sequence is shown in Table 1. Since our aim was to maximize the information from each locus developed for screening, only sequences with lengths > 450 bp were selected for further development. Approximately 25–30% of these sequences were discounted due to close matches to the mRNA, EST and mitochondrial sequences obtained from GenBank or A. cariniformis genomic resources (Table 2). The number of sequences for which we were able to generate suitable primer pairs is also shown in Table 2.

#### 3.2 Characteristics of the evaluated anonymous loci

Twelve of the 15 primer pairs tested in G lazelli amplified successfully, while only six amplified for both this taxon and Gehyra multilata. Of the four loci that sequenced successfully, one was not variable and one was polymorphic for a large (> 500 bp) indel,
A TP synthase mtDNA 808 14 (6) 14 (6) 0.1851 (0.0833)
A2 ANL 529 42 13 0.02649

The use of low coverage genome surveys has revolutionized descriptive statistics for loci compared in the study problems caused by repetitive elements by libraries. Although we have also utilized known repeat libraries a problem after screening against GenBank and known repeat approach targets low copy number sequences which, due to the for phylogenetic, phylogeographic and population studies. Our exploit these genomic resources to develop markers suitable 2011), and we have now developed a workflow to further microsatellite development in non-model organisms (Gardner et al., 2005) and the number of haplotypes recovered (n) are shown for each locus. Figures for the sea snakes include separate analyses for the overall sea snake crown group and the nested Hydrophis subgroup (in parentheses). Nucleotide diversity (\(\mu\)) was calculated using DnaSP version 5.10.01 (Librado and Rozas, 2009). Loci developed in this study are emphasized in ‘bold’. making it unsuitable as a phylogenetic marker. Of the 17 primer pairs developed from H. spinalis, 12 amplified successfully in this taxon and three other taxa spanning the sea snake crown group. Eight of these loci were polymorphic across hydrophines; however, five showed evidence of paralog co-amplification based on fixed site heterozygosity across all individuals and were discarded. Details of the loci that amplified in all species and showed no evidence of paralogs are shown in Table 3. For both reptile groups, the anonymous loci exhibited comparable or greater nuclear diversity than the other nuclear markers examined for that group.

4 DISCUSSION

The use of low coverage genome surveys has revolutionized microsatellite development in non-model organisms (Gardner et al., 2011), and we have now developed a workflow to further exploit these genomic resources to develop markers suitable for phylogenetic, phylogeographic and population studies. Our approach targets low copy number sequences which, due to the nature of the data and the methodology we have chosen, are likely to be from non-coding regions of the genome. Although these regions are likely to accumulate mutations relatively quickly, they are also rich in repetitive elements.

Thomson et al. (2008) found that repetitive elements still posed a problem after screening against GenBank and known repeat libraries. Although we have also utilized known repeat libraries in our methodology, our approach further minimizes potential problems caused by repetitive elements by de novo determination and exclusion of sequences that are repetitive within the dataset. This effectively removes species-specific SINES and other high copy number elements not present in repeat libraries or in public databases. Paralogous sequences are more difficult to detect but removing loci with sites that are heterozygous across all sequenced individuals has been demonstrated as a practical solution (Jennings and Edwards, 2005; Rosenblum et al., 2007), and this is the approach we used for the sea snake analysis.

Even though we tested a small number of loci, successful amplification in the reference taxa (80% and 71% for G. lazelli and H. spinalis, respectively) is comparable to amplification success reported in other studies (65% in Rosenblum et al., 2007; 76% in Thomson et al., 2008). During the screening process, the attrition rate of markers was high, but the reasons for attrition varied, primarily based on the phylogenetic span of trialled taxa. In the Gehyra study, 50% of the markers did not amplify in the outgroup and only one of the four sequenced products was invariant. In contrast, within the more recently diverged Hydrophis group, 33% of the loci tested were invariant and a further 42% showed evidence of paralogs.

Nevertheless, the relatively high attrition rate of loci during screening is more than offset by the large number of sequenced fragments suitable for screening. If a larger pool of potential loci are desired, increasing the size range of fragments considered or altering primer design parameters will increase the yield of potential markers.

ACKNOWLEDGEMENTS

The authors thank David Adelson and Julian Schwerdt for useful discussion regarding bioinformatic techniques and software and Stephen Donnellan and three anonymous reviewers for helpful comments on this article.

Funding: Australian Research Council Discovery Project grant to K.L.S.; M.J.S. was supported by an Australian Biological Resources Study National Taxonomy Research Grant (207-43) to S. Donnellan and M. Hutchison.

Conflict of Interest: none declared.

REFERENCES

T. Bertozzi et al.


