BIOINFORMATICS DISCOVERY NOTE Vol. 28 no. 14 2012, pages 1807–1810
doi:10.1093/bioinformatics/bts284

Data and text mining Advance Access publication May 11, 2012

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

Terry Bertozzi[1,2,*], Kate L. Sanders[3], Mark J. Sistrom[4] and Michael G. Gardner[5,6]

[1]Evolutionary Biology Unit, South Australian Museum, North Terrace, Adelaide, SA 5000, Australia, [2]School of Molecular and Biomedical Science, North Terrace, Adelaide, SA 5000, Australia, [3]Ecology, Evolution and Landscape Science, University of Adelaide, North Terrace, Adelaide, SA 5000, Australia, [4]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520-8106, USA, [5]School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia and [6]Australian Centre for Evolutionary Biology and Biodiversity, School of Earth and Environmental Science, University of Adelaide, Adelaide, SA 5000, Australia

Associate Editor: Jonathan Wren

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 multilocus 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

[*]To whom correspondence should be addressed.

of 358 bp) and 149 075 reads for *H. spiralis* (average length of 345 bp). The sequences are available from the Dryad Digital Repository doi:10.5061/dryad.f1cb2.

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_lmer_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) 'repeatmaskerlibraries-20090604' 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, 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 *Gekko* to compare to the *G. lazelli* reads. An additional BLAST database constructed from the squamate *Anolis carolinensis* genome (AnoCar 1.0, February 2007, Broad Institute) was also used to screen each dataset. Finally, matches to mitochondrial sequences from *Bungarus fasciatus* (Elapidae: GenBank accession NC_011393) and *Gekko gecko* (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 $<10^{-4}$, 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.samuseum.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.

2.2 Evaluation of anonymous nuclear loci

For all taxa examined, genomic DNA was extracted from frozen and ethanol preserved tissue using a Puregene DNA isolation kit (Gentra Systems, Minneapolis, MN, USA) following the manufacturer's protocol for DNA purification from solid tissue. All loci examined were amplified using either AmpliTaq Gold DNA polymerase (Applied Biosystems) or HotMaster *Taq* polymerase (PerkinElmer) following standard PCR protocols and visualization of double-stranded amplification products on 1.5% agarose gels. Purification of PCR products and cycle-sequencing using the BigDye Terminator v3.1 cycle-sequencing kit (Applied Biosystems) was outsourced to the Australian Genome Research Facility Ltd. (AGRF) in Brisbane, Australia.

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

Taxon	Contiguous sequence length (bp)		
	≥250	≥350	≥450
Gehyra lazelli	29 054	17 873	7957
Hydrophis spiralis	32 085	19 901	7733

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

Taxon	Sequences ≥450 bp	Sequences ≥450 bp after filtering	PCR product >400 bp	PCR product >500 bp
Gehyra lazelli	7957	5560	3874	245
Hydrophis spiralis	7733	5822	4398	76

A subset of 15 loci for *Gehyra* and 17 for hydrophiine sea snakes were assayed for amplification success and polymorphism within each group. Initial screening in *Gehyra* used *G. multilata* and *G. lazelli* which show ~15% divergence over 1200 bp of ND2 and ~5% for the protein-coding gene PRLR (Townsend *et al*., 2008).

Sea snake primers were tested for cross-amplification using four taxa that span the crown hydrophiine sea snake radiation (*H. spiralis*, *Enhydrina schistosa*, *Ephalophis greyi* and *Aipysurus apraefrontalis*). Ten additional crown group taxa were added per locus to calculate the polymorphism statistics presented in Table 3; the maximum corrected (HKY) pairwise divergence among these taxa was ~18% divergence over 808 bp of the ATP synthase gene and ~1.5% divergence at the PRLR locus. Polymorphism statistics were re-calculated for a reduced set of recently diverged *Hydrophis* subgroup species, which show maximum pairwise divergence of only 10% for the ATP synthase gene and 0.3% for the PRLR locus.

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. carolinensis* 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 mutilata*. Of the four loci that sequenced successfully, one was not variable and one was polymorphic for a large (>500 bp) indel,

Table 3. Descriptive statistics for loci compared in the study

Locus	Marker type	*L*	*n*	*h*	*µ*
Gehyra					
A1	**ANL**	**658**	**42**	**34**	**0.0612**
A2	**ANL**	**529**	**42**	**13**	**0.02649**
H3	Coding	442	100	30	0.035
PRLR	Coding	526	103	20	0.0269
MC1R	Coding	608	34	19	0.04308
RAG1	Coding	756	24	22	0.01388
ND2	mtDNA	1049	123	110	0.22
Sea snakes					
G1888	**ANL**	**439**	**14 (7)**	**8 (4)**	**0.0124 (0.0021)**
G1894	**ANL**	**445**	**14 (7)**	**12 (5)**	**0.0162 (0.0071)**
G1914	**ANL**	**494**	**14 (7)**	**13 (6)**	**0.0088 (0.0021)**
PRLR	Coding	561	14 (5)	7 (2)	0.0156 (0.0009)
cmos	Coding	915	14 (6)	11 (6)	0.0067 (0.0038)
RAG1	Coding	1002	14 (7)	12 (5)	0.0047 (0.0017)
ODC	Intron	382	6 (5)	2 (2)	0.0028 (0.0011)
ATP synthase	mtDNA	808	14 (6)	14 (6)	0.1851 (0.0833)
CYTB	mtDNA	1112	14 (7)	14 (7)	0.1471 (0.0788)

Sequence length (*L*) includes alignment gaps. The number of individuals sequenced (*n*) and the number of haplotypes recovered (*h*) 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 (*µ*) 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. spiralis*, 12 amplified successfully in this taxon and three other taxa spanning the sea snake crown group. Eight of these loci were polymorphic across hydrophiines; 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. spiralis*, 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. Hutchinson.

Conflict of Interest: none declared.

REFERENCES

Abdelkrim,J. *et al.* (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *Biotechniques*, **46**, 185–192.

Allentoft,M.E. *et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *Biotechniques*, **46**, 195–200.

Brugmans,B. *et al.* (2003) A new and versatile method for the successful conversion of AFLP markers into simple single locus markers. *Nucleic Acids Res.*, **31**, e55–e55.

Gardner,M.G. *et al.* (2011) Rise of the machines, recommendations for ecologists using next generation sequencing for microsatellite development. *Mol. Ecol. Resour.*, **11**, 1093–1101.

Jennings,W.B. and Edwards,S.V. (2005) Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. *Evolution*, **59**, 2033–2047.

Jurka,J. *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogent. Genome Res.*, **110**, 462–467.

Karl,S.A. and Avise,J.C. (1993) PCR-based assays of Mendelian polymorphisms from anonymous single-copy nuclear DNA: techniques and applications for population genetics. *Mol. Biol. Evol.*, **10**, 342–361.

Lee,J.Y. and Edwards,S.V. (2008) Divergence across Australia's carpentarian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution*, **62**, 3117–3134.

Librado,P. and Rozas,J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Lowe,A.J. *et al.* (1998) Identification and characterization of nuclear, cleaved amplified polymorphic sequence (CAPS) loci in *Irvingia gabonensis* and *I. wombolu*, indigenous fruit trees of west and central Africa. *Mol. Ecol.*, **7**, 1771–1788.

Price,A.L. *et al.* (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics*, **21**, i351–i358.

Rasmussen,D.A. and Noor,M.A.F. (2009) What can you do with 0.1x genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). *BMC Genomics*, **10**, 382.

Rosenblum,E.B. *et al.* (2007) Anonymous nuclear markers for the eastern fence lizard, *Sceloporus undulatus*. *Mol. Ecol. Notes*, **7**, 113–116.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.

Smit,A.F.A. *et al.* (1996–2010) RepeatMasker Open-3.0. Available at http://www.repeatmasker.org (last accessed date November 11, 2011).

Thomson,R.C. *et al*. (2008) Developing markers for multilocus phylogenetics in nonmodel organisms: a test case with turtles. *Mol. Phylogen. Evol.*, **49**, 514–525.

Thomson,R.C. *et al*. (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. *Mol. Ecol.*, **19**, 2184–2195.

Townsend,T.M. *et al.* (2008) Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. *Mol. Phylogenet. Evol.*, **47**,129–142.