Journal of Heredity 2011:102(1):125–129
doi:10.1093/jhered/esq099
Advance Access publication September 1, 2010

Rapid Microsatellite Development for Water Striders by Next-Generation Sequencing

Jen C. Perry and Locke Rowe

From the Department of Ecology and Evolutionary Biology, University of Toronto, Toronto ON M5S 3B2 (Perry and Rowe); and the Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, Toronto, Canada (Perry and Rowe).

Address correspondence to Jen C. Perry at the address above, or e-mail: jen.perry@utoronto.ca.

Water striders have become a model system for studies of sexual conflict and coevolution, but progress is currently limited by a lack of genetic resources. Next-generation sequencing technologies offer the potential for rapid and cost-effective development of molecular markers and hold particular promise for model organisms in ecology for which no reference genome exists. We used Roche 454 sequencing of genomic DNA to identify microsatellite loci for the water strider *Gerris incognitus*. A modest sequencing volume generated 182 912 reads, of which 30 820 (16.8%) contained microsatellite repeats. We selected 23 loci for primer development, based on criteria that maximized the likelihood of amplifying polymorphic loci, and tested them in *G. incognitus* and the related species *G. buenoi*. Of the 16 amplifying loci, 10 yielded reliable amplification and detectable polymorphism, with an average of 6.1 alleles per locus (range: 2–12). These markers should facilitate new avenues of study, including postcopulatory sexual selection, population genetic structure, phylogeography, and sexual coevolution, for a key taxon in studies of mating conflict. The current study demonstrates an effective method for microsatellite development and shows that light sequencing of genomic DNA can provide numerous and highly variable markers.

Key words: *microsatellites, 454 sequencing, sexual conflict, water striders*

The past several years have seen the remarkably rapid development of next-generation sequencing technologies (Morozova and Marra 2008). Moreover, there has recently been a surge of interest in the application of these new technologies to topics in evolutionary ecology (Ellegren 2008; Hudson 2008; Rokas and Abbott 2009; Lerner and Fleischer 2010; Wheat 2010), in large part because there is the potential to address long-outstanding questions in innovative ways. In particular, a promising application with broad utility is the prospect of rapidly identifying molecular markers, including microsatellites and single nucleotide polymorphisms, from sequence data (Table 1).

In this respect, Roche 454 pyrosequencing (R454) holds great promise, particularly for ecological model organisms (by which we mean focal species within evolutionary ecology). R454 provides read lengths that are often long enough to contain genetic information of interest within a single read. With new titanium-based chemistry, R454 can now (summer 2010) generate 100 Mbp of sequence data per run, with read lengths reaching 400–500 bp and expected to reach 1000 bp in the near future. Thus, there is increasing likelihood that a single read will contain microsatellite repeats along with suitable flanking regions of unique sequence. Other approaches to mining molecular markers in silico have also been successful (e.g., deriving markers from a draft genome, Ferrucho et al. 2009; mining existing expressed sequence tag libraries, Pan et al. 2010). However, generating new markers from R454 data offers several advantages over both traditional library–based approaches and other in silico methods. Compared with traditional approaches, deriving genetic markers via R454 is faster, less costly, and less technically demanding. Sample preparation can be as simple as extracting genomic DNA, and there is an abundance of user-friendly software available for repeat detection (reviewed by Sharma et al. 2007; Merkel and Gemmell 2008). These features make marker detection easily accessible both for laboratories with modest experience in molecular biology and for organisms with few existing genetic resources. Another advantage is the wealth of genetic information that R454 provides, yielding the possibility of future value for the data.

We used R454 pyrosequencing to identify microsatellites markers in the water strider *Gerris incognitus*, an ecological model organism for studies in sexual conflict and mating behavior (Rowe et al. 1994). We report the successful development of 10 polymorphic primers, test their cross-amplification in a related species, and survey other studies to

Table 1 Characteristics of studies reporting microsatellites developed from 454 pyrosequencing, including the current study

Study	Organism	No. of individuals sequenced	Sample sequenced	Mbp	No. of reads	Mean read length (bp)	% reads containing microsatellites	Development time
Present study	Water strider (*Gerris incognitus*)	5	¼ plate	61.5	182 912	369	16.8%	3 weeks
Castoe et al. (2010)	Copperhead snake (*Agkistrodon contortrix*)	1	≥3/8 plate	27	128 773	215	11.3%	~9 days
Abdelkrim et al. (2009)	Blue duck (*Hymenolaimus malacorhynchos*)	Multiple	1/16 plate	4.1	17 215	243	1.3%	2 weeks
Allentoft et al. (2009)	Heavy-footed moa (*Pachyornis elephantopus*)[a]	1	¼ plate		79 796	112	0.24%	
Santana et al. (2009)[b]	Pine pathogen fungus (*Fusarium circinatum*)	Multiple	Single lane	1.67	8.644		97%	2 months
Santana et al. (2009)[b]	Pine-damaging wasp (*Sirex noctilio*)	Multiple	Single run	1.47	7016		25% (FIASCO enrichment); 50% (ISSR enrichment)	
Santana et al. (2009)[b]	Nematode parasite (*Deladenus siricidicola*)	Multiple	Single run	1.22	6388		40%	
Rasmussen and Noor (2009)	Scuttle fly (*Megaselia scalaris*)	Multiple	¼ plate		129 080	231		
Vanpé et al. (2009)	Short-beaked echidna (*Tachyglossus aculeatus*)		12 runs[c]		885 433		0.84%	

Empty cells indicate that the paper did not report this information.

[a] Extinct.

[b] Sequencing performed on DNA enriched for microsatellites.

[c] Existing 454 sequence data from 12 runs were accessed for this study.

examine typical success in microsatellite development using R454. Although the successful development of microsatellites via R454 has been reported for several vertebrates, there have been only 2 examples in insects and none from the Hemiptera (Table 1). Given that insect taxa vary substantially in the ease of microsatellite isolation (Zhang 2004; Arthofer et al. 2007) and that some genetic techniques have proved difficult to apply to water striders (Damgaard 2008), this study provides a test case for a new insect order.

Materials and Methods

Sample Preparation and 454 Pyrosequencing

Gerris incognitus water striders were collected from Laughlin Lake (British Columbia, Canada; 48.948365, −123.502679) and stored in 95% ethanol. Voucher specimens have been deposited at the Royal Ontario Museum (Toronto, Canada). DNA was extracted from strider leg muscles following the Gentra Puregene Tissue Protocol (Qiagen, Valencia, CA). We assessed DNA quality by spectrophotometric absorbance and electrophoresis on a 1% agarose gel.

We pooled 5 μg of DNA from 5 apterous individuals for a quarter plate of sequencing on a Roche 454 GS FLX sequencer with titanium chemistry (McGill University and Génome Québec Innovation Centre, Montréal QC).

Selecting Markers

We used the program Exact Tandem Repeat Analyzer 1.0 (E-TRA, Karaca et al. 2005; available from ftp://ftp.akdeniz.edu.tr/) to identify reads and contigs containing microsatellite repeats. Our criteria were a minimum of 5 repeats for simple motifs or 3 repeats for complex or imperfect repeats and a motif length from 2 to 10 bp. Of the reads identified, we selected a subset that had both a minimum of 8 repeats and a maximum homopolymer of 3 nucleotides within the repeat motif. We checked these reads for suitable primers using Primer3 (v. 0.4.0; Rozen and Skaletsky 2000) and used the following criteria to identify loci with a good likelihood of reliable amplification: GC content > 30%, final product length 150–300 bp, primer length 18–25 bp, a maximum homopolymer of 5 nucleotides within the primer sequence, and maximum difference in melting temperature between paired primers of 3 °C. Of the loci that met these criteria, we synthesized 23 for further testing.

Testing Markers

During primer synthesis, we labeled the reverse primers with an additional sequence that was complementary to a separate M13 sequence, which we tagged with a fluorescent dye (FAM, NED, PET, or VIC; Schuelke 2000). We evaluated each locus using DNA samples from 10 individuals in polymerase chain reactions (PCRs) using these cycling conditions: an initial denaturing for 4 min at 94 °C; 34 cycles of 94 °C for 15 s, the primer-specific annealing temperature for 25 s, and 72 °C for 25 s; and a 10 minute final extension at 72 °C to maximize adenylation (thermal cycler: Bio-Rad MyCycler v. 1.065, Hercules, CA). The annealing

temperature for the final 8 of the 34 cycles was decreased to the annealing temperature of the M13 primer (50 °C; Schuelke 2000). The total reaction volume was 12.5 μl and consisted of 1× ThermoPol buffer (New England BioLabs, Beverly, MA), 0.2 mM of each dNTP, 0.5 units of *Taq* DNA polymerase, 3.0 pmoles each of the forward primer and the dye-labeled M13 oligo, 1.0 pmoles of the reverse primer (Schuelke 2000), and ~100 ng of DNA. We calculated an initial annealing temperature for each primer pair based on the T_m for each primer and adjusted accordingly if this initial temperature generated multiple bands or did not amplify with all samples. Each PCR was run with a negative and positive control.

The PCR products were visualized by electrophoresis on a 2% agarose gel. Some loci did not initially amplify for all samples or yielded double or faint bands. In these cases, we adjusted PCR conditions, and if the outcome remained inconsistent, we excluded the locus from further testing. For the remaining loci, we conducted reactions with an additional 16 DNA samples and genotyped the total of 26 samples on an ABI 3730 xL DNA Analyzer (Applied Biosystems, Foster City, CA) through the Centre for Applied Genomics (Toronto, Canada). The samples were multiplexed for genotyping by pooling samples tagged with different dyes within a well.

We assessed the reliability of the primers by repeating the amplification and genotyping for 27 of the samples (10.4%). We tested the potential for cross-amplification using 5 DNA samples from a congener, *G. buenoi*. To assess whether the loci were located near a coding sequence, we used each locus sequence in basic alignment search tool (BLAST) searches (v. 2.2.23; Zhang et al. 2000) against the pea aphid (*Acyrthosiphon pisum*) and fruit fly (*Drosophila melanogaster*) genomes, using an E value cutoff of 10^{-5}.

We tested for evidence of deviations from Hardy–Weinberg equilibrium, genotyping errors, null alleles, and large allele dropout using MICRO-CHECKER v. 2.2.3 (van Oosterhout et al. 2004). We tested for linkage disequilibrium between each pair of loci using GENEPOP 4.0.10 (Rousset 2008).

Results

Pyrosequencing

The R454 sequencing run yielded 61.4 Mbp of data and 182 912 reads of average length 369 bp, with 30 820 reads (16.8%) containing microsatellite repeats (Table 1). In total, 23 318 reads contained a simple perfect repeat, whereas 11 449 contained a compound or imperfect repeat; 3947 contained more than one repeated sequence. The GC content for all reads was 32.08%, within a range commonly observed for insects (e.g., Archak et al. 2007).

Within the reads containing simple repeats, dinucleotides dominated (88.5%; Supplementary Table S1). The largest motif class—decanucleotides—was present at very low frequency (0.02%). A sizeable fraction of the simple repeats contained a large number of repeats, with at least 24 repeats occurring in hundreds of dinucleotides and several trinucleotides and tetranucleotides (Supplementary Table S1). Among the nonsimple repeats, the majority were imperfect repeats (85.6%) and a smaller fraction were compound repeats (8.6%) or extended compound repeats (5.8%).

Microsatellite Development

We selected 23 microsatellite loci to synthesize and test. Of these, 16 amplified with most of the samples tested (judging by the presence of bands on an agarose gel). Of these 16, 5 gave multiple bands or faint bands and were not considered further. Of the remaining 11, 10 could be easily genotyped. All 10 displayed polymorphism (Table 2). The average number of alleles per locus was 6.1. Expected heterozygosities ranged from 0.13 to 0.81 (= 0.58, standard deviation [SD] = 0.25) and observed heterozygosities from 0.14 to 0.75 (= 0.49, SD = 0.20).

There was no evidence of large allele dropout or genotyping errors due to stutter peaks. One locus (Gi38) showed homozygote excess (deviation from Hardy–Weinberg equilibrium, $P < 0.01$; Table 2) and therefore evidence of a null allele with an estimated frequency of 11.5%. There was no evidence of linkage disequilibrium (GENEPOP: P values for each pairwise population comparison >0.05). Genotyping was quite consistent: all 27 samples tested twice yielded the same genotype. BLAST searches did not detect matches between these 10 microsatellite sequences and 2 insect genomes (all E values $>10^{-5}$), thus providing no evidence that the microsatellites are located near coding regions.

Eight loci amplified with *G. buenoi* DNA (Gi20, Gi31, Gi36, Gi38, Gi43, Gi56, Gi57, Gi71), whereas 2 did not (Gi69, Gi72).

Discussion

This study joins a handful of others, now across a broad range of taxa (Table 1), that have successfully employed random pyrosequencing to generate numerous polymorphic microsatellites. Our study is the first of these to take advantage of R454 titanium chemistry. This new chemistry allows longer read lengths, reflected in the longer average read length in our data compared with previous studies (Table 1). Along with increasing the probability of detecting microsatellite repeats and suitable primers within a read, longer reads increase the likelihood of detecting loci with more repeats, which are expected to be more polymorphic (e.g., Lai and Sun 2003). Comparing the number of loci detected across studies is difficult because genomes vary substantially in their frequency of microsatellites; however, the number of loci detected here was similar to or higher than other studies (Table 1). In all these studies, the number of microsatellites detected is most likely inflated by the occurrence of multiple reads covering the same sequence. The frequency with which this occurs will be influenced by genome coverage and is expected to be low for many genome sizes and read lengths (see Castoe et al. 2010).

Table 2 Characteristics of microsatellite loci reported in this study

Locus repeat motif	Primer sequence (5′-3′)	T_A (°C)	Sample size	N alleles	Allele sizes (bp)	H_o	H_e	GenBank accession no.
Gi20 $(TAA)(AT)_n(TAA)(AT)_n(AT)_n(G)(TA)_nC(TA)_n(AT)_n$	F: ACCCCAGAACATTGCATCTC; R: AGGGTTCGGACACAAAAACA	51	21	2	180–182	0.19	0.23	HM770904
Gi31 AGAAT	F: TTCCAGTTTGTAATTTTCCGTCT; R: TGCCAGAACATCAAAACATCA	51	22	7	212–304	0.57	0.74	HM770905
Gi36 ACAA	F: TGGTGCATTGTCTTCGTGTT; R: GGGGGAAACCTATGCCTACA	51	22	7	186–210	0.75	0.81	HM770906
Gi38 ACAT	F: TCGAAGGTTTGTTTGTTGAATG; R: TGTATCCTGTGTATGTATTGCTGA	51	22	9	155–224	0.54	0.81	HM770907
Gi43 $(ATAC)_n(ATAT)(ATAC)_n$	F: CCATAAAGATTCCGACGCTAA; R: TGTGTTTGAGAAAGGGTTAATAATG	53	22	5	206–236	0.61	0.70	HM770908
Gi56 TAG	F: CTCCTGTGTCTCCCTTTTGC; R: GGCAAAGTCTACCCATTCCA	53	20	8	196–223	0.55	0.71	HM770909
Gi57 CA	F: CACGTCGAGAGTGAGCTGAA; R: GGCAATTCAATGGGAGACTG	53	22	5	223–242	0.30	0.40	HM770910
Gi69 CAT	F: TCAATTCATTCAAAGGTTTCTTG; R: AACAGAGAATCGAGCCCAAA	50	20	12	217–281	0.65	0.78	HM770911
Gi71 AGT	F: GGTGGGGAAGGAAGAACATT; R: GATGTGTGAAGTGGCTGGAA	53	21	4	227–246	0.58	0.53	HM770912
Gi72 ACT	F: TGCAAGGTTCAAACAAGTGG; R: AACACAAGCCCAAAGAATGG	50	21	2	239–242	0.14	0.13	HM770913

The advantages of generating microsatellites by pyrosequencing were realized in this study. We were able to locate loci rapidly and with decreased costs (ca. $4500) compared with outsourcing a traditional library–based approach ($10 000 to $15 000); moreover, costs of R454 continue to drop, along with improvements in read lengths which mean that less coverage will be required to achieve a similar microsatellite output. A further advantage to R454 is that many loci are detected, allowing the targeted selection of loci most likely to amplify and exhibit polymorphism. A relatively high proportion of the 23 loci we tested amplified successfully, and all those that amplified proved to be polymorphic (Table 2).

There are, however, potential difficulties to marker detection by pyrosequencing. One risk is the failure to detect any useable loci. Although published studies have reported success (Table 1), it is possible that unsuccessful attempts are not reported; still, the fact that even very low coverage can generate dozens of loci (e.g., Rasmussen and Noor 2009) suggests that this risk may be low. A related risk is the failure to detect loci with sufficient polymorphism (e.g., when sequencing individuals from populations that have experienced a recent bottleneck, Wheat 2010); however, a recent study detected polymorphic loci for even an endangered species with reduced genetic diversity (Abdelkrim et al. 2009). Another concern is the potential for slippage errors in homopolymeric sequences with R454 (Hudson 2008). However, often the number of reads containing microsatellite repeats will be large enough to allow the selection of loci without extended homopolymers.

Obtaining genetic markers through pyrosequencing offers exciting possibilities for further applications, including the identification of markers for extinct organisms (e.g., Allentoft et al. 2009) and developing microsatellites for multiple species in a single sequencing run (see Binladen et al. 2007). It may also be possible to use R454 to identify microsatellites that are linked to coding sequence in a reference genome, by submitting R454 data to bulk BLAST searches. Although microsatellites are rarely found in coding sequence, they more regularly occur in intron sequences (Li et al. 2002), and the large numbers of microsatellites from R454 increase the odds of a match. Finding such loci would yield genetic markers potentially linked to functional loci of interest and may provide information on the location of the microsatellite within the genome.

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Funding

Natural Sciences and Engineering Research Council (NSERC) and the Canada Research Chairs program to L.R.; scholarships from NSERC and the Government of Ontario (Dr F. M. Hill Scholarship in Science and Technology) to J.P.

Acknowledgments

We thank A. Bruce, A. Cutter, M. Dixon, B. Fraser, W. Thomas, C. Weadick, and 2 anonymous reviewers for advice and comments on the manuscript, and W. Cole for laboratory assistance.

References

Abdelkrim J, Robertson BC, Stanton J-AL, Gemmell NJ. 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques. 46:185–191.

Allentoft ME, Schuster SC, Holdaway RN, Hale ML, McLay E, Oskam C, Gilbert MTP, Spencer P, Willerslev E, Bunce M. 2009. Identification of microsatellites from an extinct moa species using highthroughput (454) sequence data. BioTechniques. 46:195–200.

Archak S, Meduri E, Kumar PS, Nagaraju J. 2007. InSatDb: a microsatellite database of fully sequenced insect genomes. Nucleic Acids Res. 35:D36–D39.

Arthofer W, Schlick-Steiner BC, Steiner FM, Avtzis DN, Crozier RH, Stauffer C. 2007. Lessons from a beetle and an ant: coping with taxon-dependent differences in microsatellite development success. J Mol Evol. 65:304–307.

Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE. 2:e197.

Castoe TA, Poole AW, Gu W, de Koning APJ, Daza JM, Smith EN, Pollock DD. 2010. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. Mol Ecol Resour. 10:341–347.

Damgaard J. 2008. MtDNA diversity and species phylogeny of western Palaearctic members of the *Gerris lacustris* group (Hemiptera–Heteroptera: Gerridae) with implications for "DNA barcoding" of water striders. Insect Syst Evol. 39:107–120.

Ellegren H. 2008. Sequencing goes 454 and takes large-scale genomics into the wild. Mol Ecol. 17:1629–1631.

Ferrucho RL, Zala M, Zhang Z, Cubeta MA, Garcia-Dominguez C, Ceresini PC. 2009. Highly polymorphic *in silico*-derived microsatellite loci in the potato-infecting fungal pathogen *Rhizoctonia solani* anastomosis group 3 from the Colombian Andes. Mol Ecol Resour. 9:1013–1016.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour. 8:3–17.

Karaca M, Bilgen M, Onus AN, Ince AG, Elmasulu SY. 2005. Exact tandem repeats analyzer (E-TRA): a new program for DNA sequence mining. J Genetics. 84:49–54.

Lai Y, Sun F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 20:2123–2131.

Lerner HRL, Fleischer RC. 2010. Prospects for the use of next-generation sequencing methods in ornithology. The Auk. 127:4–15.

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 11:2453–2465.

Merkel A, Gemmell N. 2008. Detecting short tandem repeats from genome data: opening the software black box. Briefings Bioinform. 9:355–366.

Morozova O, Marra MA. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics. 92:255–264.

Pan L, Xia Q, Quan Z, Liu H, Ke W, Ding Y. 2010. Development of novel EST–SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. J Hered. 101:71–82.

Rasmussen DA, Noor MAF. 2009. What can you do with 0.1× genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). BMC Genomics. 10:382–390.

Rokas A, Abbott P. 2009. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 24:192–200.

Rousset F. 2008. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour. 8:103–106.

Rowe L, Arnqvist G, Sih A, Krupa JJ. 1994. Sexual conflict and the evolutionary ecology of mating patterns: water striders as a model system. Trends Ecol Evol. 9:289–293.

Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, Krawetz S, editors. Miseners. Bioinformatics methods and protocols: methods in molecular biology. Totowa (NJ): Humana Press. p. 365–386.

Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, Wingfield MJ, Wingfield BD. 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. BioTechniques. 46:217–223.

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol. 18:233–234.

Sharma PC, Grover A, Kahl G. 2007. Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 25:490–498.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P. 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes. 4:535–538.

Vanpé C, Buschiazzo E, Abdelkrim J, Morrow G, Nicol SC, Gemmell NJ. 2009. Development of microsatellite markers for the short-beaked echidna using three different approaches. Aus J Zool. 57:219–224.

Wheat CW. 2010. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica. 138:433–451.

Zhang D-X. 2004. Lepidopteran microsatellite DNA: redundant but promising. Trends Ecol Evol. 19:507–509.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7:203–214.

Received April 30, 2010; Revised August 10, 2010; Accepted August 11, 2010

Corresponding Editor: Lacey Knowles