

Multiplexed microsatellite recovery using massively parallel sequencing

T. N. JENNINGS,* B. J. KNAUS,* T. D. MULLINS,† S. M. HAIG† and R. C. CRONN*

*Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, OR 97331, USA,

†Forest and Rangeland Ecosystem Science Center, US Geological Survey, 3200 SW Jefferson Way, Corvallis, OR 97331, USA

Abstract

Conservation and management of natural populations requires accurate and inexpensive genotyping methods. Traditional microsatellite, or simple sequence repeat (SSR), marker analysis remains a popular genotyping method because of the comparatively low cost of marker development, ease of analysis and high power of genotype discrimination. With the availability of massively parallel sequencing (MPS), it is now possible to sequence microsatellite-enriched genomic libraries in multiplex pools. To test this approach, we prepared seven microsatellite-enriched, barcoded genomic libraries from diverse taxa (two conifer trees, five birds) and sequenced these on one lane of the Illumina Genome Analyzer using paired-end 80-bp reads. In this experiment, we screened 6.1 million sequences and identified 356 958 unique microreads that contained di- or trinucleotide microsatellites. Examination of four species shows that our conversion rate from raw sequences to polymorphic markers compares favourably to Sanger- and 454-based methods. The advantage of multiplexed MPS is that the staggering capacity of modern microread sequencing is spread across many libraries; this reduces sample preparation and sequencing costs to less than \$400 (USD) per species. This price is sufficiently low that microsatellite libraries could be prepared and sequenced for all 1373 organisms listed as 'threatened' and 'endangered' in the United States for under \$0.5 M (USD).

Keywords: genomic library, Illumina, microsatellite, multiplex massively parallel sequencing

Received 18 February 2011; revision received 29 April 2011; accepted 5 May 2011

Introduction

Molecular characterization of natural populations of conservation and management concern depend on accurate and inexpensive genotyping methods. There has been significant improvement in new genotyping approaches in the past few years (Brumfield *et al.* 2003; Morin *et al.* 2004; Miller *et al.* 2007), but traditional microsatellite analysis remains one of the most popular genotyping methods because of the low cost of development, ease of analysis and high exclusion probabilities, even in the limited gene pools of elite crop germplasm (e.g., Hamblin *et al.* 2007; Narum *et al.* 2008; Glover *et al.* 2010). These codominant marker systems are arrays of short repetitive motifs [i.e., dinucleotide patterns such as (AC)_n] that are distributed throughout the genome. Generation of polymorphism at these sites is believed to be due largely to slippage of the template during replication, and this process results in an increase or decrease in the number of elements in the array (Ellegren 2004). The high frequency

at which mutations occur at these sites produces the high level of polymorphism required for population genetic analysis. Owing to these properties, microsatellites are widely used to make inferences on population structure, connectivity, heterozygosity and parentage (Pemberton *et al.* 2009; Ouborg *et al.* 2010).

In contrast to newer SNP-based methods, the identification and screening of microsatellites is not so much technically or computationally demanding as it is laborious. The recent development of massively parallel sequencing (MPS) has simplified and streamlined the traditional process of microsatellite development (e.g., Glenn & Schable 2005) in two key ways. First, bacterial library construction is eliminated because isolation and amplification of each DNA fragment takes place directly on the sequencing platform. Second, MPS sequences a large number of DNA fragments simultaneously, up to hundreds of millions per sample per run (Mardis 2008; Metzker 2010). As MPS is most commonly used for sequencing full genomes or subgenomic partitions (such as transcriptomes), microsatellite development from these platforms most commonly involves scanning for microsatellite repeats in total genomic (e.g., Abdelkrim

Correspondence: Richard C. Cronn, Fax: +1 541 750 7329, E-mail: rcronn@fs.fed.us

et al. 2009; Allentoft *et al.* 2009; Castoe *et al.* 2009) or transcriptomic (e.g., Mikheyev *et al.* 2010; Parchman *et al.* 2010) pools, although sequencing of microsatellite-enriched genomic fractions has also been performed (Santana *et al.* 2009; Martin *et al.* 2010; Malausa *et al.* 2011). To date, MPS-based identification has focused primarily on the 'long-read' pyrosequencing chemistry offered by the Roche/454 sequencer, because this platform offered the comparatively long reads required to identify microsatellite motifs and sufficient flanking DNA for amplification primer design (Santana *et al.* 2009; Boomer & Stow 2010; Csencsics *et al.* 2010; Martin *et al.* 2010; Saarinen & Austin 2010; Malausa *et al.* 2011).

The recent and dramatic increase in read length in 'microread' sequencers now makes these platforms viable alternatives to pyrosequencing for identifying microsatellite-containing sequences, particularly when paired-end reads are obtained. In their initial development, microread sequencers (e.g., Solexa/Illumina; SOLiD) yielded sequence reads in the range of 18–25 bp (Mardis 2008). While they produced millions of microreads, these sequences were too short to span microsatellites, making it impossible to produce locus-specific contigs. In their current form, sequencing read lengths of 80 bp are routine for the Illumina Genome Analyzer, and the recent release of newer chemistry now makes it possible to obtain individual sequence reads over 150 bp in length. When combined with paired-end sequencing (e.g., sequencing from each end of a DNA molecule; Bentley *et al.* 2008), molecules over 300 bp in length can now be sequenced to completion, making it possible to sequence complete microsatellites and flanking DNA for primer design and—given the output of these instruments—from an exceedingly large number of microsatellites per run.

These improvements in sequencing have already facilitated the generation of microsatellite markers in a number of plant and animal taxa for which no markers were previously available and at much reduced time and expense compared with traditional methods (Santana *et al.* 2009; Boomer & Stow 2010; Csencsics *et al.* 2010; Martin *et al.* 2010; Saarinen & Austin 2010; Malausa *et al.* 2011). Widespread adoption of these methods is certain to lower the cost of microsatellite development, allowing population and conservation biologists to develop microsatellites for any organism. In this article, we show how multiplexed MPS ('MMPS'; Cronn *et al.* 2008) can add to this already efficient approach by permitting simultaneous enrichment and sequencing of multiple microsatellite libraries. Through the use of multiplexing adapters, libraries developed from different species are coded with a molecular 'barcode', then pooled, selectively hybridized and sequenced using long paired-end sequences (in this example, 80 bp per end for 160 bp total). Once sequenced, individual sequences are parsed by barcode,

and microsatellites are screened within individual microreads. Here, we show how MMPS was applied to recover microsatellites from two species of conifers and five species of birds and how Illumina-based microsatellite recovery compares with other published 'next-generation' examples.

Materials and methods

DNA isolation and library construction

Total genomic DNA was isolated from conifer needles or megagametophyte tissue (Port-Orford Cedar (POC), *Chamaecyparis lawsoniana*; Alaska Yellow Cedar (AYC), *Callitropsis nootkatensis*) (Willyard *et al.* 2007); blood or pectoralis tissue of avian species (Yellow Rail (YER), *Coturnicops noveboracensis*; Pfrimer's Conure (CON), *Pyrrhura pfrimeri*; Gull-billed Tern, *Gelochelidon nilotica*; Red-Cockaded Woodpecker (RCW), *Picooides borealis*; Micronesian Kingfisher, *Todiramphus cinnamominus*), as described previously (Haig *et al.* 1994). Genomic DNAs were sheared to approximately 300–400 bp with a Bio-Ruptor sonicator (Diagenode Inc., Denville, NJ, USA) using an ultrasonic wave power setting of 320 W for 30 min. Once sheared, the DNAs were converted to Illumina libraries as previously described (Cronn *et al.* 2008). Library construction followed the recommended Illumina protocols, with the exception that internally 'bar-coded' adapters (Craig *et al.* 2008; Cronn *et al.* 2008) were substituted for standard paired-end adapters. Adapters used in this effort included three index nucleotides and a 3'-terminal 'T' required for adapter ligation. All DNA-cleaning steps used Agencourt AMPure paramagnetic beads (Beckman Coulter Genomics), as described by Quail *et al.* (2008). DNA from all steps (input; library) was quantified using the Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA).

Hybridization-based enrichment and multiplexed massively parallel sequencing

A detailed protocol for this experimental approach is provided in the supplementary information. For each microsatellite hybridization selection, we added 0.5–1 µg of barcoded DNA library to a hybridization reaction (40 µL) that contained the following reagents: 6× SSC, 5 mM EDTA, 0.1% SDS, 1× Denhardt's reagent, 2 µg of lambda DNA and 100 pmol of each biotinylated microsatellite probe. For this experiment, we included three microsatellite probes, specifically 5'-[Biotin](CT)₁₅, 5'-[Biotin](GT)₁₅, and 5'-[Biotin](CA)₁₅ (Integrated DNA Technologies, Coralville, IA, USA). Dinucleotide probes were chosen for this particular experiment, although probes containing any repeat motif of interest could be

used. This mixture was heated to 95 °C for 10 min and then incubated for 64–72 h at 50 °C for hybridization.

Prior to hybrid capture, 50 µL of streptavidin-coated paramagnetic beads (MagnaSphere, Promega Inc., Madison, WI, USA) were equilibrated in 200 µL of 2× casein bead blocking buffer (Sigma Chemical Co., St. Louis, MO, USA) for 2 h. These were rinsed two times in 200 µL of T₁₀N₁₀E₁ (room temperature) and then resuspended in 200 µL of high-salt binding buffer (1 M NaCl/10 mM Tris-HCl (pH 8.0)/1 mM EDTA). The 40 µL hybridization reaction was added to the magnetic bead/binding buffer suspension, mixed well and then incubated for 30 min. Beads were magnetically separated, the supernatant was removed, and beads were washed twice with 200 µL of 1× SSC/0.1% SDS for 15 min at 50 °C. Following the second wash, beads were magnetically separated, the supernatant discarded and the beads washed with 200 µL of 0.5× SSC/0.1% SDS at 50 °C for 10 min, followed by a final capture and wash with 200 µL of 0.1× SSC/0.1% SDS at 50 °C for 10 min. Following this wash, beads were magnetically separated, the supernatant discarded, and the enriched library was eluted from the streptavidin beads with 50 µL of water at 80 °C for 10 min.

After elution, the microsatellite-enriched library was given a final PCR enrichment to increase the DNA concentration. Library PCRs included 25 µL Phusion Flash 2× premix (Finnzymes Oy), 10 µL of hybridization-enriched DNA library, 25 pmol each of Illumina paired-end PCR primers 1.1 and 2.1 and water to 50 µL. Cycling conditions were as follows: 98 °C for 30 s, followed by 8–12 cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s and ending with 72 °C for 5 min. Cycles were kept to a minimum to prevent biasing of library and accumulation of identical reads. Gel-based size selection of amplified DNA allowed us to isolate a library that ranged between 300 and 400 bp in size, with correctly adapted molecules containing 126 bp of adapter sequence (65 bp from PCR primer 1.0; 61 bp from PCR primer 2.0; Bentley *et al.* 2008) and inserts of 174 bp to 274 bp. As 80 bp paired-end reads sequenced a total of 160 bp from each molecule, this left approximately 14–114 bp of the centre of each microread unsequenced (Fig. 1). Aliquots of libraries were run on electrophoretic gels to verify size distributions, and the libraries were cleaned and quantitated as noted previously.

Indexed libraries were pooled in approximately equimolar concentrations and then submitted for paired-end 80 bp sequencing at the Oregon State University Center for Gene Research and Biocomputing (CGRB; <http://www.cgrb.oregonstate.edu/>). Pooled libraries were loaded onto one lane of an Illumina Genome Analyzer II at 5 pM concentration. Cluster generation and sequencing used Illumina version 3.0 reagents, and image acquisition and base calling used the Illumina pipeline version 1.5.

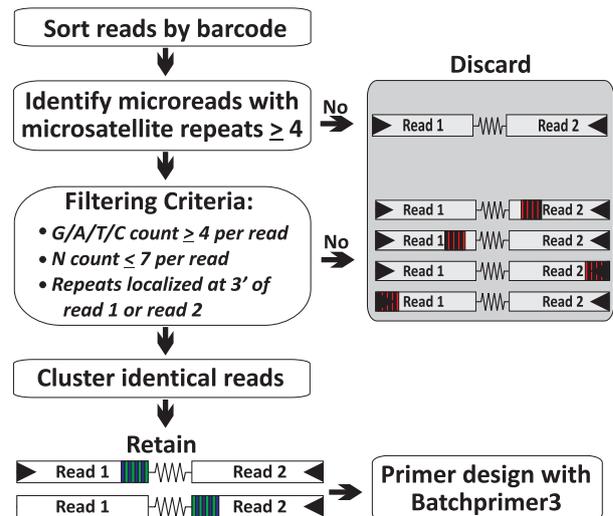


Fig. 1 Flow chart for informatic microsatellite identification by microread-based multiplexed massively parallel sequencing. Green vertical hashes indicate the location of targeted microsatellites, while red vertical hashes indicate microsatellites that did not fit selection criteria.

Informatic identification of microsatellites and potential microsatellite primer sites

Microreads were sorted by barcode (4 bp) using a custom perl script (bcsort, <http://brianknaus.com/software/srtoolbox/shortread.html>). Sorted reads were then searched for the motifs used for enrichment [(AC)_n and (AG)_n], as well as the unenriched dinucleotides (AT)_n and (GC)_n, and all trinucleotide motifs. Microreads containing at least four perfect repeats, with each nucleotide represented at least four times, and fewer than eight ambiguous bases were retained for screening (Fig. 1). Paired-end microsatellite-containing reads were joined into a single sequence by concatenating read 1 and the reverse complement of read 2, separated by 50 Ns (to identify the break between microreads). Identical and nearly identical sequences are expected in microsatellite enrichment methods (Glenn & Schable 2005), so we filtered redundant sequences (identity ≥95%) to a single unique microread using the program cd-hit-454 (Niu *et al.* 2010). Finally, we applied a stringent filter to identify microreads with microsatellites located near the centre of the molecule, such that the 3'-end of read 1 or 2 terminated in an SSR (ssr_sorter_pe.pl, <http://brianknaus.com/software/srtoolbox/shortread.html>). This stringent filter ensures that microsatellite repeats are located near the centre of a DNA molecule, and it yields the largest possible flanking sequences for subsequent primer design. The filtered SSR-containing sequences were then evaluated with BatchPrimer3 (You *et al.* 2008) to identify PCR primer sites. We used the default settings for all but

the following parameters: product length (min = 100 bp, max = 200 bp); primer length (min = 17 bp, max = 25 bp, optimum = 19 bp); melting temperature (min = 48 °C, max = 63 °C, optimum = 54 °C); number of primers per sequence = 1.

Microsatellite primer validation

From our list of potential primers, we chose a subset to screen on unpedigreed populations from four organisms, including two conifer species (POC and AYC) and two bird species (YER and CON). Individuals for each species were selected from a single population, as opposed to across a broad geographic range. Primers were designed from various microsatellite motifs. Microsatellite amplification products were initially labelled using polymerase incorporation of fluorescent dCTP (Perkin Elmer, Boston, MA, USA) using the manufacturer's recommendations and then screened at the Oregon State University CGRB using capillary electrophoresis (ABI3730 with POP7 polymer, 50 cm capillaries; Applied Biosystems Inc., Foster City, CA, USA). Once polymorphism was validated, primers were re-synthesized using 5'-end-labelled fluorophores (6-FAM, HEX). Electrophoretograms were analysed using GeneMapper ver. 4.0 (Applied Biosystems Inc.). For our purposes, PCR amplification of a microsatellite was considered successful if a band of the expected size was observed on gel electrophoresis, even if two or more bands were amplified. A microsatellite locus was considered successful if the resulting electrophoretograms showed at least two alleles per locus, but never more than two alleles from an individual, in a screen of 10 individuals from one population.

Results

Of the 8 389 278 sequences returned from the Illumina pipeline, 6 121 628 sequences contained the expected barcodes for our species, and 1 305 168 sequences were identified that contained dinucleotide (1 279 468) and trinucleotide (25 700) microsatellites (Table 1). Individual libraries included between 107 902 and 298 035 microsatellite-containing sequences each, with a relative microsatellite abundance ranging from 15% to 31% per library. We screened for two unenriched dinucleotide motifs [(AT)_n and (GC)_n] and found that they ranged in abundance from 0.8% to 1.6% in these libraries (Fig. 2). While we did not make a direct comparison of unenriched versus enriched microsatellite frequencies, the comparatively high abundance of (AC)_n and (AG)_n relative to unenriched dinucleotide motifs offer a rough approximation of the hybridization enrichment success under these conditions. Enrichment is an important source of differentiation for recovered motifs, but certain

Table 1 Microsatellite ('SSR') enrichment using low-stringency hybridization and seven-plex MMPS. 'N/A' indicates primers were not tested for these species or that the value is not applicable

| Common name | Port-Orford cedar | Yellow cedar | Yellow rail | Pfimer's conure | Gull-billed tern | Red-cockaded woodpecker | Micronesian kingfisher | Lane totals (average) |
|--------------------------------------|---------------------------------|----------------------------------|-----------------------------------|-------------------------|------------------------------|--------------------------|---------------------------------|-----------------------|
| Scientific name | <i>Chamaecyparis lawsoniana</i> | <i>Callitropsis nootkatensis</i> | <i>Coturnicops noveboracensis</i> | <i>Pyrhura pfrimeri</i> | <i>Gelochelidon nilotica</i> | <i>Picoites borealis</i> | <i>Todiramphus cinnamominus</i> | |
| Family | Cupressaceae | Cupressaceae | Rallidae | Psittacidae | Sternidae | Picidae | Halcyonidae | |
| Total reads | 2 020 681 | 489 625 | 412 682 | 439 854 | 1 264 228 | 863 635 | 630 923 | 6 121 628 (874.5) |
| Total sequences containing SSRs | 298 035 | 136 785 | 107 902 | 137 168 | 245 703 | 242 641 | 137 168 | 1 305 168 (186 559) |
| % SSR abundance | 15 | 28 | 26 | 31 | 19 | 28 | 22 | N/A (24) |
| Unique sequences containing SSRs | 92 703 | 34 394 | 52 532 | 29 423 | 42 349 | 72 619 | 32 938 | 356 958 (50 994) |
| Filtered dinucleotide SSR sequences | 2758 | 742 | 2144 | 1349 | 1616 | 1671 | 1105 | 11 385 (1626) |
| Filtered trinucleotide SSR sequences | 117 | 23 | 38 | 8 | 24 | 35 | 20 | 265 (38) |
| Dinucleotide Batch Primer3 sequences | 2720 | 734 | 2100 | 1303 | 1549 | 1598 | 1098 | 11 102 (1586) |
| Primer pairs tested | 144 | 96 | 30 | 50 | N/A | N/A | N/A | 320 (80) |
| Loci amplified | 134 | 92 | 24 | 41 | N/A | N/A | N/A | 294 (73.5) |
| Loci scorable | 15 | 13 | 24 | 27 | N/A | N/A | N/A | 79 (20) |
| Alleles per locus (average) | 10 | 8 | 6 | 8 | N/A | N/A | N/A | N/A (8) |
| Amplicon length, min/max | 97/394 | 147/443 | 84/228 | 169/311 | N/A | N/A | N/A | N/A |

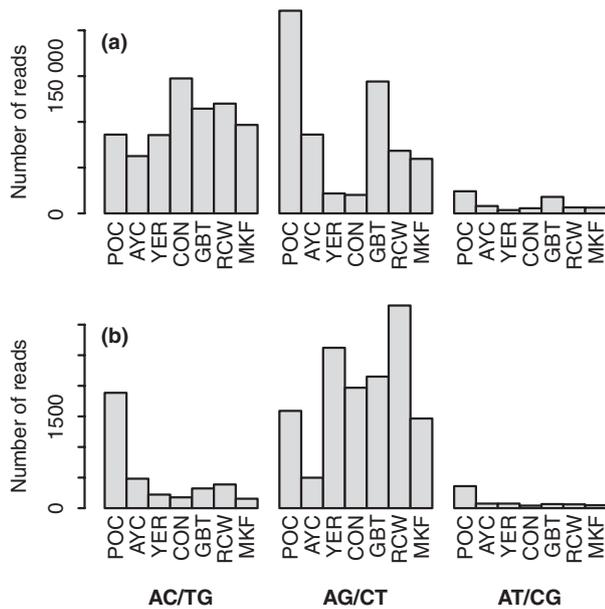


Fig. 2 Counts of microsatellite-containing sequences. (a) Number of microreads containing a microsatellite motif anywhere in the read, prior to filtering for redundancy or complexity. (b) Number of unique microreads containing a microsatellite after stringent filtering for position of repeat and sequence complexity in the flanking region (see text). Libraries were enriched for $(AC)_n$ and $(AG)_n$ motifs. Because the Illumina sequencing process does not preserve strandedness, targeted motifs and their complement (e.g., AC and TG) are pooled.

motifs are likely to show very different relative abundance in individual genomes, such as CT/AG motifs that are abundant in our selected conifer samples but rare in our selected bird samples (Supporting information).

The process of filtering identical to nearly identical repeats reduced the total number of available microsatellite-containing sequences to 356 958 (344 840 dinucleotide, 12 118 trinucleotide). As our goal was to identify microsatellites localized near the centre of a DNA molecule, we imposed a highly stringent filter that retained only those sequences where one of the two reads terminated on a SSR greater than or equal to four repeats in length (Fig. 1). This stringent filter removed an average of 96.8% of the repeat-containing sequences, leaving a pool of 11 385 unique, filtered dinucleotide SSR-containing sequences per species (range = 742–2758; Table S1; Supporting information). Nonredundant trinucleotide SSR-containing sequences were also detected (265 total), and these showed partial homology to the dinucleotide probes used in enrichment (Table S1; Supporting information). These high-quality sequences were analysed using BatchPrimer3, where we found that 97% of the dinucleotide microsatellite-containing microreads were suitable for microsatellite PCR primer design.

We chose to design microsatellite primers for four species: AYC, POC, YER and CON. We synthesized a total of 320 primers, 294 of which showed success on initial amplification tests (92%) with 8 individuals from each species (Table 1). Based on this initial success, we re-synthesized 118 primers with 5' fluorophores and screened these primers against a collection of 10 individuals for one population per species. Of these primers, 79 showed banding patterns consistent with a diploid locus/allele model, with conifers showing a much lower average success rate (12%) than birds (78%). Across loci and species, these 79 polymorphic markers averaged 8 alleles/locus, with an average amplicon length of 208 bp (range 97–443; Table 1). Detailed descriptions of these markers will be provided in separate reports for each species.

Discussion

Microsatellites occupy an important niche in the genotyping world. As a 'mature' technology, microsatellites are increasingly outshined by newer techniques, such as single-nucleotide addition/SNP assays (Brumfield *et al.* 2003; Morin *et al.* 2004), or genotyping by high-throughput sequencing (Miller *et al.* 2007; Craig *et al.* 2008). These newer approaches offer higher throughput and broader genomic coverage, but at a substantial cost in terms of community effort for marker development, curation and analysis; critically, they are not superior to microsatellites in all applications (Hamblin *et al.* 2007; Haasl & Payseur 2011). In contrast, microsatellite technology has been widely used for decades, so the behaviour of the markers, methods for data analysis and the limits of the data are established and well known (Ellegren 2004; Pemberton *et al.* 2009; Ouborg *et al.* 2010). While microsatellites do not scale as efficiently as SNP approaches for genome-wide coverage, the power of microsatellites lies in their high per-marker variability, a property that confers high exclusion probabilities (e.g., paternity analysis), high likelihood of segregation in closely related pedigrees (e.g., linkage mapping populations) and high per-marker information content in diversity studies (Hamblin *et al.* 2007; Haasl & Payseur 2011). For these reasons, microsatellites will continue to play an important role in providing the genetic information needed to study, manage and conserve genetic resources in nonmodel and wild species.

An important advantage offered by multiplexed MPS approaches relative to nonmultiplexed methods (Santana *et al.* 2009; Boomer & Stow 2010; Csencsics *et al.* 2010; Martin *et al.* 2010; Saarinen & Austin 2010) is that the enormous capacity of microread sequencing can be spread across many SSR-enriched libraries simultaneously; this adds substantially to the throughput, efficiency and cost savings of an already-powerful, but

Table 2 Comparison of studies using next-generation sequencing to develop microsatellites

| Published Study | Organism | Source | Platform* (geometry) | Reads (avg L, bp) | Repeats Screened | Nonredundant SSRs identified | Estimated cost, SSR/\$† | Batch primer3 loci |
|------------------------------|--|---------------------------|------------------------------|----------------------|--|---------------------------------|----------------------------|-----------------------|
| Abdelkrim <i>et al.</i> 2009 | Blue Duck (<i>Hymenolaimus malacorrhynchos</i>) | Total genomic DNA | Roche 454 (1/16 PTP) | 17 215 (243) | ≥7 repeats for di-, tri- and tetranucleotides | 231 | 0.18 | 24 |
| Allentoft <i>et al.</i> 2009 | Extinct Moa (<i>Aves: Dinornithiformes</i>) | Total genomic DNA | Roche 454 (1/4 PTP) | 79 796 (112) | Motif and length not specified | 195 | 0.09 | N.R.‡ |
| Castoe <i>et al.</i> 2009 | Copperhead snake (<i>Agkistrodon contortrix</i>) | Total genomic DNA | Roche 454 (1/4 + 1/8 PTP) | 128 773 (215) | ≥6 repeats for di-, ≥ 4 tri- and ≥ 3 tetranucle otides | 14 612 | 1.89 | 4564 |
| Santana <i>et al.</i> 2009 | Pitch canker fungus (<i>Fusarium circinatum</i>) | ISSR-enriched DNA | Roche 454 (1/16 PTP) | 8644 (193) | ≥5 repeats for di-, tri- and tetranucleotides | 1644 | 1.31 | 231 |
| Santana <i>et al.</i> 2009 | Sirex woodwasp (<i>Sirex noctilio</i>) | FIASCO-enriched DNA | Roche 454 (1/16 PTP) | 7016 (210) | ≥5 repeats for di-, tri- and tetranucleotides | 975 | 0.78 | 496 |
| Santana <i>et al.</i> 2009 | <i>Deladenus siricidicola</i> | FIASCO-enriched DNA | Roche 454 (1/16 PTP) | 6388 (188) | ≥5 repeats for di-, tri- and tetranucleotides | 421 | 0.34 | 296 |
| Boomer & Stow 2010 | Gummy shark (<i>Mustelus antarcticus</i>) | Total genomic DNA | Roche 454 (1/8 PTP) | 86 756 (N.R.) | ≥6 repeats for di-, ≥ 4 tri- and ≥ 4 tetranucleotides | 4362 | 3.48 | N.R.‡ |
| Csencsics <i>et al.</i> 2010 | Dwarf bulrush (<i>Typha minima</i>) | Total genomic DNA | Roche 454 (1/8 PTP) | 76 692 (341) | ≥8 repeats for di-, ≥ 9 tri- and ≥ 6 tetranucleotides | 307 | 0.24 | 98 |
| Mikheyev <i>et al.</i> 2010 | Checkerspot Butterfly (<i>Euphydryas editha</i>) | Total cDNA | Roche 454 (1 PTP) | 864 056 (N.R.) | ≥8 repeats for di-, tri- and tetranucleotides | 92 | 0.01 | 72 |
| Parchman <i>et al.</i> 2010 | Lodgepole pine (<i>Pinus contorta</i>) | Total cDNA | Roche 454 (1 PTP) | 586 732 (306) | ≥3 repeats for di-, tri- and tetranucleotides | 15 084 | 2.10 | 4020 |
| Saarinien & Austin 2010 | Okaloosa Darter (<i>Etheostoma okaloosae</i>) | Total genomic DNA | Roche 454 (1/8 PTP) | 82 463 (352) | ≥6 repeats for di-, ≥ 4 tri- and ≥ 4 tetranucleotides | 7420 | 5.91 | 1600 |
| Malaua <i>et al.</i> 2011 | 8 animals, 2 fungi, 3 plants | Hybridized genomic DNA | Roche 454 (1/8 + 1/4 PTP) | 281 506 (211) | ≥5 repeats for di-, tri- and tetranucleotides | 28 336 | 6.41 | 6915 |
| This study | 2 conifers, 5 birds, multiplex | Hybridized genomic DNA | Illumina (1 lane) | 6 121 628 (160) | ≥4 repeats for di- and trinucleotides | 356 958 | 132.2 | 11 650 |

*Geometry refers to either the proportion of a picotitre plate (PTP) used for 454-based sequencing or the number of flow cell lanes used for Illumina-based sequencing.

†Cost per microsatellite assume these prices (in \$US): Library construction costs at \$100 each for genomic DNA, \$200 for cDNA; 454 sequencing costs at \$7000 for 1 picotitre plate (PTP), \$3500 for 1/2 PTP, \$1965 for 1/4 PTP, \$1155 for 1/8 PTP or less; Illumina Genomic Analyzer sequencing costs at \$2000 for one lane of paired-end 80-bp sequence.

‡N.R. = information not reported in original publication.

nonmultiplexed, approach. This approach was recently adopted for the development of multiple microsatellite libraries using the 454 GS-FLX pyrosequencer (Malausa *et al.* 2011), and similar efficiencies and cost savings were realized on this platform. It is worth noting, however, that Illumina-based MPS has evolved during the course of our study from a platform that could deliver 10–20 million 80-bp paired-end reads per lane, to current instruments that deliver over 100 million 100-bp paired-reads per lane. These sequencing lengths are sufficient to identify microsatellite motifs and sufficient flanking DNA for primer design, and at an order of magnitude higher sample multiplexing than what we attempted here.

Another advantage of our approach is that the extremely high capacity of microread sequencers can be used to provide estimates on the relative abundance of a repeat in the library. The large and highly repetitive genomes of conifers are known to be an important obstacle in microsatellite conversion success, as microsatellite motifs are abundant in the highest copy fractions in these genomes (Elsik & Williams 2001). Indeed, the low rate of scoring success we experienced in conifers relative to birds (12% vs. 78%, respectively) may be directly related to the comparatively large size of conifer genomes (approximately 12 Gbp for *Chamaecyparis* and *Callitropsis*; Hizume *et al.* 2001) and the relatively compact nature of bird genomes (approximately 1.5 Gbp; Organ *et al.* 2007). While we did not incorporate repeat abundance information in our primer selection strategy, adding additional filters to select for unique to moderately rare (presumably low copy) repeats, or exclude highly abundant repeats, would likely improve the recovery of amplifiable and interpretable microsatellites.

The main disadvantage of the microread MPS approach is that careful attention must be made to the insert size of the library. If the inserts are larger than the sum of the two read lengths (as is the case in our study), inserts will be only partially sequenced. This complicates the analysis, as many microsatellite-containing microreads may go undetected if the repeat is in the middle of an insert. In addition, incompletely sequenced molecules are less well characterized than they would be with longer sequencing technologies, and sequence motifs that are undesirable for microsatellite analysis (e.g., compound repeats, single-nucleotide repeats) may not be detected. While it is advisable to control insert sizes so that they are closely matched to the total read length, our results show that large numbers of useful microsatellites can still be identified when molecules are incompletely sequenced. As read lengths continue to improve on microread sequencers, this drawback will diminish in importance. Finally, we note that a commonly cited drawback of microread sequencing—specifically, the high depth requirement for sequence assembly or SNP

detection (Craig *et al.* 2008)—does not apply in the case of microsatellite identification, as every microread passing quality criteria serves as a finished sequence.

To emphasize the comparative efficiency of different MPS methods, we estimated the approximate cost required to clone, sequence and identify one nonredundant microsatellite-containing sequence from a diversity of published organisms using nonmultiplexed and multiplexed sequencing on the 454/Roche, or multiplexed sequencing on a microread sequencer (such as the Illumina Genome Analyzer; Table 2). Our microread-based approach produced a minimum of 20 times more microsatellites per unit currency as compared with methods based on the 454/Roche platform (Table 2). At the multiplexing level used in this study (seven-plex), the estimated cost for library construction and sequencing is less than USD \$400 per library. At this price, community efforts to develop and sequence microsatellite libraries could proceed quickly, making genetic markers a universally available commodity for organisms of management concern. For example, microsatellite libraries could be prepared and sequenced in the manner described here for the 1373 ‘threatened’ and ‘endangered’ species in the United States (http://ecos.fws.gov/tess_public/pub/boxScore.jsp) for approximately USD \$0.52 M. As noted, this cost is certainly an overestimate, as the recent 10-fold increase in sequence output demonstrated by microread platforms favours a many-fold higher multiplexing level than the seven-plex described here.

Acknowledgements

The authors thank Jennifer Swanson, Katie Alderman, Kelly McDonald and Mari Carrillo for laboratory assistance, and Mark Dasenko and Chris Sullivan for sequencing and core facility computing support. Scott Kolpak, Dave D’Amore and Paul Hennon provided population samples for conifers. Funding for this research was provided by the U.S. Forest Service Forest Health and Protection STDP program (for conifer research), the Pacific Northwest Research Station, the U.S. Forest Service Dorena Genetic Resource Center and the U.S. Geological Survey. Any use of trade, product or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

- Abdelkrim J, Robertson BC, Stanton J-AL, Gemmill 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.
- Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
- Boomer J, Stow A (2010) Rapid isolation of the first set of polymorphic microsatellite loci from the Australian gummy shark, *Mustelus*

- antarcticus*, and their utility across divergent shark taxa. *Conservation Genetics Resources*, **2**, 393–395.
- Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.
- Castoe TA, Poole AW, Gu W *et al.* (2009) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources*, **10**, 341–347.
- Craig DW, Pearson JV, Szeling S *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*, **5**, 887–893.
- Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**, e122.
- Csencsics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered Dwarf Bulrush (*Typha minima*) using next-generation sequencing technology. *Journal of Heredity*, **101**, 789–793.
- Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.
- Elsik CG, Williams CG (2001) Low-copy microsatellite recovery from a conifer genome. *Theoretical and Applied Genetics*, **103**, 1189–1195.
- Glenn TC, Schable NA (2005) Isolating microsatellite DNA loci. In: *Methods in Enzymology 395, Molecular Evolution: Producing the Biochemical Data, Part B* (eds Zimmer EA & Roalson EH), pp. 202–222. Academic Press, San Diego, CA. ISBN: 9780121828004.
- Glover KA, Hansen MM, Lien S, Als TD, Hoyheim B, Skaala O (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, **11**, 2.
- Haas RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, **106**, 158–171.
- Haig SM, Rhymer JM, Heckel DG (1994) Population differentiation in randomly amplified polymorphic DNA of red-cockaded woodpeckers *Picoides borealis*. *Molecular Ecology*, **3**, 581–595.
- Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS ONE*, **2**, e1367.
- Hizume M, Kondo T, Shibata F, Ishizuka R (2001) Flow cytometric determination of genome size in the Taxodiaceae, Cupressaceae *sensu stricto* and Sciadopityaceae. *Cytologia*, **66**, 307–311.
- Malausa T, Gilles A, MeglÉCz E *et al.* (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources*, doi: 10.1111/j.1755-0998.2011.02992.x.
- Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.
- Martin JF, Pech N, MeglÉcz E *et al.* (2010) Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **11**, 560.
- Metzker ML (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**, 31–46.
- Mikheyev AS, Vo T, Wee B, Singer MC, Parmesan C (2010) Rapid microsatellite isolation from a butterfly by de novo transcriptome sequencing: Performance and a comparison with AFLP-derived distances. *PLoS ONE*, **5**, e11212.
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.
- Morin PA, Luikart G, Wayne RK, SNP Workshop Group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.
- Narum SR, Banks M, Beacham TD *et al.* (2008) Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, **17**, 3464–3477.
- Niu BF, Fu LM, Sun SL, Li WZ (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, **11**, 187. doi:10.1186/1471-2105-11-187.
- Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature*, **446**, 180–184.
- Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics*, **26**, 177–187.
- Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.
- Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA (2009) Sequence determinants of human microsatellite variability. *BMC Genomics*, **10**, 612.
- Quail MA, Kozarewa I, Smith F *et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, **5**, 1005–1010.
- Saarinen EV, Austin JD (2010) When technology meets conservation: increased microsatellite marker production using 454 genome sequencing on the endangered Okaloosa darter (*Etheostoma okaloosae*). *Journal of Heredity*, **101**, 784–788.
- Santana QC, Coetzee MPA, Steenkamp ET *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217–223.
- Willyard A, Syring J, Gernandt DS, Liston A, Cronn R (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Molecular Biology and Evolution*, **24**, 620–620.
- You FM, Huo NX, Gu YQ *et al.* (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, **9**, 253.

Data Accessibility

DNA sequences: DNA sequence files were deposited at the European Bioinformatics Institute sequence read archive under accession number ERP000647 (<http://www.ebi.ac.uk/ena/data/view/ERP000647>).

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Experimental protocol used for isolating microsatellites using multiplexed massively parallel sequencing.

Table S2 Frequencies of dinucleotide- and trinucleotide-containing sequences identified from 11 650 unique paired-end 80 bp sequences.

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.