MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) **12**, 448–455 doi: 10.1111/j.1755-0998.2011.03102.x

Cross-species transferability of SSR loci developed from transciptome sequencing in lodgepole pine

MARK R. LESSER,*† THOMAS L. PARCHMAN† and C. ALEX BUERKLE*†
Program in Ecology, University of Wyoming, Laramie, WY 82071, USA, †Department of Botany, University of Wyoming, Laramie, WY 82071, USA

Abstract

With the advent of next generation sequencing technologies, transcriptome level sequence collections are arising as prominent resources for the discovery of gene-based molecular markers. In a previous study more than 15 000 simple sequence repeats (SSRs) in expressed sequence tag (EST) sequences resulting from 454 pyrosequencing of *Pinus contorta* cDNA were identified. From these we developed PCR primers for approximately 4000 candidate SSRs. Here, we tested 184 of these SSRs for successful amplification across *P. contorta* and eight other pine species and examined patterns of polymorphism and allelic variability for a subset of these SSRs. Cross-species transferability was high, with high percentages of loci producing PCR products in all species tested. In addition, 50% of the loci we screened across panels of individuals from three of these species were polymorphic and allelically diverse. We examined levels of diversity in a subset of these SSRs by collecting genotypic data across several populations of *Pinus ponderosa* in northern Wyoming. Our results indicate the utility of mining pyrosequenced EST collections for gene-based SSRs and provide a source of molecular markers that should bolster evolutionary genetic investigations across the genus *Pinus*.

Keywords: EST, *Pinus contorta*, *Pinus ponderosa*, 454 pyrosequencing, SSR

Received 30 August 2011; revision received 21 October 2011; accepted 9 November 2011

Introduction

Simple sequence repeats (SSRs), or microsatellites, are short, tandemly repeated DNA motifs that occur throughout the genomes of most organisms in both coding and noncoding regions (Toth *et al.* 2000). Because of slippage errors that occur in DNA replication of these sequences (Schlötterer 1998), they have high mutation rates, typically resulting in many allelic states at individual loci in natural populations (Ellegren 2004; Selkoe & Toonen 2006). Despite the usefulness of SSRs, their development can be difficult because of high species specificity, and laborious and expensive laboratory methods (Zane *et al.* 2002).

The recent surge in the availability of genomic-level sequence collections has made the detection and characterization of SSR loci easily achievable with simple bioinformatics approaches (Ellis & Burke 2007). Repeat regions with virtually any desirable qualities can be located in genomic-level sequence collections, and primers for polymerase chain reaction (PCR) can be readily developed. Because whole genome-level sequence data are still scarce for most groups of organisms, large collections of expressed sequence tags (ESTs) currently offer the most promising source of information for SSR discovery and characterization (Gupta *et al.* 2003; Ellis & Burke 2007). ESTs are sequences representing expressed genes and are a good source for the development of molecular markers for several reasons. SSRs occurring in ESTs typically have higher amplification rates and cross-species transferability owing to the relatively conserved nature of protein-coding DNA sequences (Hempel & Peakall 2003; Barbara *et al.* 2007; Ellis & Burke 2007). Gene-based SSRs are also desirable for many applications, as they allow investigations of variability to focus on the functional gene space of organisms (Gutierrez *et al.* 2005; Slate *et al.* 2007). A potential drawback of using EST-based SSRs is that they may experience selection. However, the majority of SSRs even in genic regions will not experience selection (Woodhead *et al.* 2005), and it remains unclear how much of an issue this might be in comparison with nongene-based markers (Ellis & Burke 2007). A second potential issue is that EST-based SSRs can have lower

Correspondence: Mark R. Lesser, Fax: 307 766 2851;
E-mail: mlesser@uwyo.edu

polymorphism rates than SSRs in noncoding regions (Gupta *et al.* 2003; Chagne *et al.* 2004). The higher cross-species transferability and functional aspects of these markers out-weigh this drawback. EST-based SSRs have been widely applied in genetic mapping (Moccia *et al.* 2009), population genetics (Kim *et al.* 2008; Simko 2009) and population genomics approaches (Vasemagi *et al.* 2005).

Until recently, the development of EST sequence collections was laborious and cost intensive (Bouck & Vision 2007). Next generation sequencing technologies, however, now allow the sequencing of large EST collections at a fraction of the time and cost previously required (Hudson 2008; Mardis 2008; Wheat 2008). As a result such resources are rapidly growing and becoming publicly available and offer a rich source of information for the *in-silico* development of SSRs and other genetic markers. Recent studies utilizing pyrosequencing of ESTs and simple bioinformatics approaches have demonstrated the ability to rapidly detect and characterize thousands of gene-based SSRs (Novaes *et al.* 2008; Meyer *et al.* 2009; Castoe *et al.* 2010; Parchman *et al.* 2010).

Despite its ecological and economic importance, whole genome sequencing efforts and associated marker development for the genus *Pinus* have lagged behind other groups. This paucity exists in part because conifers have enormous genomes (10 000–40 000 mega-base pairs vs. 114.5 Mbp in *Arabidopsis thaliana*) containing large amounts of repetitive DNA (Guevara *et al.* 2005; Ralph *et al.* 2006), making whole genome sequencing projects difficult. Deep divergence times for many groups in the genus have also limited the cross-species transferability of markers (Echt *et al.* 1999). Consequently, the construction of EST collections offers a promising approach for providing genome level resources in pines (Ralph *et al.* 2006; Neale 2007). Substantial Sanger sequencing efforts over the last decade have produced a large collection of EST sequences for select pine species [e.g. *Pinus taeda* (loblolly pine), *Pinus pinaster* (maritime pine) and *Pinus palustris* (longleaf pine); Echt & Nelson 1997; Chagne *et al.* 2004; Liewlaksaneeyanawin *et al.* 2004; Berube *et al.* 2007], but such resources have been limited for other pines. We recently used 454 pyrosequencing to generate an EST collection for *Pinus contorta* (lodgepole pine; Parchman *et al.* 2010). More than 300 000 unique sequences were generated by this work, spanning more than 18 000 unique genes and serving as a basis for the identification of thousands of molecular markers, including SSRs (Parchman *et al.* 2010).

Here we tested for successful amplification, polymorphism and rate of cross-species transferability for 184 SSR loci characterized in Parchman *et al.* (2010). We tested these SSRs across nine species in the genus *Pinus* that have not previously benefitted from thorough marker development. We characterized SSRs with high cross-species transferability that should provide additional resources for genetic analyses across the genus *Pinus*. Finally, we quantified levels of diversity and differentiation for a subset of these loci across several populations of *Pinus ponderosa* (ponderosa pine). Our results highlight the value of next generation transcriptome resources for the characterization and development of gene-based SSRs and provide a valuable resource for population genetic studies in *Pinus*.

Materials and methods

A computer program in the Perl language was written to identify SSRs in transcriptome sequences and to identify a subset of these that reside in regions where EST sequences from *Pinus contorta* were assembled onto *Pinus taeda* unigenes (Parchman *et al.* 2010). We located di-, tri- and tetra-nucleotide SSRs with lengths <50 bp and with a minimum of five contiguous repeating units, which provided a large number of candidate SSRs (Parchman *et al.* 2010). BATCHPRIMER3 (You *et al.* 2008) was used to construct PCR primers in the flanking regions of SSRs. We designed primers that were a minimum of 12 bp long and used stringent criteria to target genetic markers of desired PCR product length and with a high probability of amplification. We created primers with a minimum GC content of 30%, a melting temperature between 52 and 62 °C, and a maximum 4 °C difference in melting temperature between primer pairs. Primers were positioned to obtain PCR products between 100 and 450 bp long. We also constrained primer construction so that the end of each primer contained a GC clamp (the last nucleotide was G or C).

We selected a subset of 184 SSRs that contained perfect di-, tri- or tetra-nucleotide repeats with a minimum length of 15 bp to test for amplification rate, polymorphism and cross-species transferability. Ninety-four of these SSRs were characterized based on assembled consensus sequences from EST sequences (contigs), and 90 were created from singleton sequences that were only represented once in the transcriptome data. Of the markers developed in contigs, 39 were characterized in contigs resulting from aligning EST sequences from *P. contorta* with a unigene set available for *P. taeda* at NCBI (http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3352, Pta.seq.all.gz file, Build #12) as described previously (Parchman *et al.* 2010). Thus, the conserved nature of the priming sites may result in higher cross-species transferability levels [See Appendix S2 (Supporting information) for information on the identity of the sequence containing each SSR, SSR motif, number of repeats, expected length of PCR product, sequence and position of primers, and annealing temperature].

We obtained needles as source of DNA for eight pine species, in addition to *P. contorta* (Table 1). We chose these species so as to include new and old world representatives of the hard pines (subgenus *Pinus*), as well as the more highly divergent soft pines (subgenus *Strobus*). Additionally these species spanned a wide range of phylogenetic distances so that transferability of markers across the entire genus could be assessed. DNA was extracted from a minimum of eight individuals of each species using a modified version of a cetyl-trimethyl ammonium bromide (CTAB) protocol (Doyle 1991). Extracted DNA was subjected to electrophoresis to confirm the presence of high molecular weight template for PCR and quantified using a Nanodrop spectrophotometer (Thermo Scientific, Inc.).

We tested for successful amplification of each SSR locus in each of the nine species. Here and throughout, we scored loci amplifiable when PCR resulted in a product in the expected size range, as detected on 1.5% agarose gels. PCR amplification reactions consisted of 50–100 ng total genomic DNA; 2 pmol of each primer; 0.5 mM each of dATP, dCTP, dGTP and dTTP; 1×PCR buffer; and 0.4 units of Taq polymerase. All PCR amplifications were performed with the following conditions: 94 °C for 5 min, followed by 32 cycles of 94 °C for 1 min, 45 °C for 1 min and 72 °C for 1 min, followed by a final extension step of 72 °C for 3 min. We subjected 18 μL of each product to electrophoresis on 1.5% agarose gels stained with ethidium bromide. Because of the large numbers of markers and individuals screened, and because we wished to design markers that amplified with ease and did not require PCR optimization, we tested each SSR locus for successful amplification on DNA from each species in only two individual PCRs. For each reaction, we scored results as positive or negative for amplification, but also recorded markers that resulted in more than one discrete PCR product.

We tested for polymorphism in 35 loci in *P. contorta*, *Pinus ponderosa* and 20 loci in *Pinus longaeva* that had high amplification rates across the surveyed species. We amplified each locus across eight individuals in each of the above species and determined allelic size and diversity by subjecting the products to electrophoresis on 4% Metaphor agarose gels (Lonza, Inc.). The precise resolution of these gels allowed us to discern heterozygous individuals and fragment size differences between individuals within a very small size range. Metaphor agarose has been shown to distinguish between alleles sizes as small as 4 bp, and to be as accurate as polyacrylamide gels (Ong *et al.* 2009). We used a 20 bp ladder to determine the size fragments. We were able to easily distinguish different size fragments and detect the presence of multiple fragments in heterozygous individuals; however, we were unable to definitively size individuals below the resolution of the 20 bp range that they fell into on the ladder. This level of precision was more than adequate for polymorphism screening, but could not be used for accurate genotyping.

From the 35 loci tested for polymorphism, we chose six that were polymorphic for *P. ponderosa* (two or more distinct size bands across the eight individuals screened) to investigate patterns of variation across natural populations. We obtained genotypes for these six markers in 1127 individuals across four populations of *P. ponderosa* occurring in north central Wyoming that are the focus of ongoing population genetic studies (M.R. Lesser, T.L. Parchman & S.T. Jackson, in preparation). For these loci, we added fluorescent dyes to primers to allow detection on a capillary DNA sequencer. PCR products were run on an ABI 3130 capillary sequencer at the Genome Center at the University of Nevada, Reno. Sizing of fragments and allele binning was carried out using the Genemapper software (ABI, Inc.). For each locus, we calculated the number of alleles, allele size range, and expected and

Table 1 Species included in this study, and the number and percentage of 184 tested simple sequence repeats that amplified

Species	Number of individuals	Sample location	Number amplified	Per cent amplified
Pinus contorta	32	Vedauwoo, WY	145	0.79
Pinus banksiana	8	Thunder Bay, Ontario, CA	129	0.70
Pinus ponderosa	32	Bighorn Basin, WY	135	0.73
Pinus palustris	8	Benton, GA	116	0.63
Pinus elliotii	8	Gainesville, FL	120	0.65
Pinus halepensis	16	Tucson, AZ	127	0.69
Pinus longaeva	8	Laramie, WY	107	0.58
Pinus edulis	8	Granite Canyon, CO	99	0.54
Pinus flexilis	16	Vedauwoo, WY	105	0.57

observed heterozygosities (H_0). We calculated estimated frequency of null alleles using the program MICROCHECKER (van Oosterhout *et al.* 2004). We calculated genetic diversity statistics and estimates of population differentiation statistics using the program Microsatellite Analyzer (Dieringer & Schlötterer 2003) and SMOGD (Crawford 2010). To adjust F_{ST} estimates for the presence of null alleles, we used the excluding null alleles (ENA) approach of Chapuis & Estoup (2007). We present estimates of F_{IS}, F_{ST} (Weir & Cockerham 1984) and Jost's D_{est} (Jost 2008) as metrics of genetic differentiation. To evaluate the informativeness of markers, polymorphic information content (PIC) was calculated following Botstein *et al.* (1980).

Results

More than 15 000 SSRs were identified in the 303 480 sequences representing contigs and singletons in the *Pinus contorta* transcriptome sequenced ESTs characterized in Parchman *et al.* (2010), giving a rate of 0.05 SSRs per nonredundant EST sequence. PCR primers were successfully constructed using BATCHPRIMER3 (You *et al.* 2008) for 4020 of these candidates (Parchman *et al.* 2010) (Appendix S1, Supporting information). Of 184 SSRs tested here, 98 consisted of di-nucleotide repeats, 75 consisted of tri-nucleotide repeats and 11 consisted of tetra-nucleotide repeats (Appendix S2, Supporting information). Among the tested SSRs, the average length of the repeat regions was 34 bp (Appendix S2, Supporting information). The targeted PCR fragments containing SSRs averaged 174 bp in length and ranged between 85 and 311 bp. Fifty-four of the sequences containing SSRs had BLAST matches to proteins in UniRef50 (Suzek *et al.* 2007) at an *E*-value threshold of 10^{-10}.

Of the 184 SSRs screened, 145 (79%) amplified readily in *P. contorta* (Table 1) in initial amplification screening. Relatively high percentages of these loci also amplified in the other species tested, with amplification rates ranging from 54% to 73% (Table 1). A small number of tested loci gave multiple bands, potentially indicating that these SSRs reside in paralogous sequences (Appendix S2, Supporting information). Perhaps as a result of large intronic regions, some loci produced bands much larger than the expected size, which were not considered further. Pairwise sequence divergence of each species from *P. contorta* was estimated based on matK cpDNA sequences described in Eckert & Hall (2006) (GenBank accession no.: AB063499, AB080921, DQ168629, AB080931, DQ168637, AB019856, AF456368, AB063505, AB063506). The success of per species SSR amplification generally decreased as per cent sequence divergence from *P. contorta* increased (Fig. 1, $P = 0.003$). A higher percentage of the SSRs designed from contig consensus sequences amplified in *P. contorta* (92.6%) than those designed from singleton sequences (65.6%, $P < 0.0001$) [see Appendix S2 (Supporting information) for SSR origin]. SSRs designed from contigs also had higher cross-species transferability than those designed from singleton sequences ($P < 0.001$). The SSRs designed in contigs resulting from assembly of *P. contorta* transcriptome sequences onto *Pinus taeda* unigenes were consistent with a higher transferability rate than other SSRs, although this was not statistically significant. Cross-species transferability was not different for di-, tri- and tetra-nucleotide repeats ($P = 0.58$) or for SSRs of different length ($P = 0.21$).

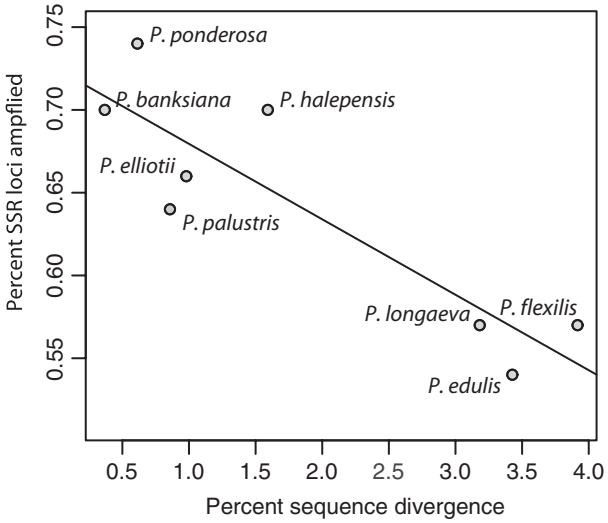

Fig. 1 Per cent successful amplification of 184 simple sequence repeats (SSRs) by per cent sequence divergence (chloroplast matK sequences) from *Pinus contorta* for eight *Pinus* species.

Of the 35 loci that were assayed for allelic variability in *P. contorta* and *Pinus ponderosa*, 18 were polymorphic in the former and 15 in the latter. Twenty loci were assayed in *Pinus longaeva* with six being polymorphic (Table 2). The number of alleles in polymorphic loci that could be definitively scored on the Metaphor agarose gels ranged from two to three across the eight individuals tested from each species (Table 2).

From the 15 polymorphic loci identified for *P. ponderosa*, six were selected for further analysis (Table 3). These six loci were amplified in 1127 individuals across four populations of *P. ponderosa*. The number of alleles at each locus ranged from 9 to 25. Observed and expected heterozygosities varied across loci, but with the exceptions of Pico3 and Pico31 were generally high. PIC scores were also high, with four of the six loci having highly informative scores (PIC > 0.50), and the other two having reasonably informative scores (0.50 > PIC > 0.25) (Botstein *et al.* 1980) (Table 3). However, all six loci showed significant deficiencies in observed heterozygosity (Table 3).

Table 2 Thirty-five simple sequence repeat loci tested for polymorphism with motif and number of repeats. Number of alleles and size range of alleles are given for eight individuals of *Pinus contorta*, *Pinus ponderosa* and *Pinus longaeva*

Locus	ID	Motif	No. Repeats	*P. contorta*		*P. ponderosa*		*P. longaeva*	
				No. alleles	Size range	No. alleles	Size range	No. alleles	Size range
Pico1	PtaS26629966	CTT	18	3	160–200	1	160–180	1	140–160
Pico2	PtaS21729441	GTG	18	2	180–220	2	180–220	x	x
Pico3	PtaS25806063	TGC	15	1	180–200	3	200–220	2	160–180
Pico7	PtaS21728471	GCA	15	2	200–220	1	200–220	1	180–220
Pico14	PtaS20933364	CAG	15	2	180–200	1	180–200	x	x
Pico15	PtaS15767430	CAT	15	1	180–200	1	160–180	1	160–180
Pico17	PtaS26641723	CCA	15	1	160–180	1	180–200	2	160–200
Pico21	PtaS15752020	GCG	21	1	200–220	2	180–200	x	x
Pico24	PtaS26014332	GAG	15	1	240–260	1	220–240	x	x
Pico25	PtaS25806179	GCA	15	2	200–240	2	200–220	1	200–220
Pico26	PtaS25587627	CTG	18	3	180–220	2	160–200	1	160–180
Pico27	PtaS25797870	GAG	18	1	180–200	2	200–240	1	180–200
Pico31	PtaS20694200	CCT	18	1	180–200	2	160–180	2	180–200
Pico37	PtaS25557090	TTG	15	1	180–200	2	180–200	1	180–200
Pico42	348	TCC	18	1	160–180	1	160–180	2	160–180
Pico43	1268	TCT	18	2	220–240	1	200–220	1	200–220
Pico44	1755	TCT	18	1	220–240	1	200–220	1	200–220
Pico49	8778	TTA	15	1	180–200	1	200–220	1	180–200
Pico50	10236	CAA	15	1	200–220	1	180–200	1	180–200
Pico54	17020	GTG	18	2	180–220	2	200–220	1	160–180
Pico57	17510	GAA	18	1	180–200	1	180–200	x	x
Pico66	Repeat-33759	CTT	18	2	160–200	1	180–200	1	160–180
Pico68	Repeat-35369	TCC	18	2	200–220	2	180–200	2	160–180
Pico71	Repeat-42294	CTT	18	2	180–220	1	180–200	x	x
Pico73	Repeat-44006	TCA	18	1	180–200	1	200–220	1	180–200
Pico85	Repeat-53073	CT	18	2	180–220	2	180–220	2	180–220
Pico104	FTBEI0F16JUPVB.1	ATT	42	2	120–180	3	120–180	x	x
Pico109	FTC1UFH02F4YR0.1	TG	34	3	160–200	3	140–180	x	x
Pico114	FTC1UFH02FFWG9.1	CT	32	1	200–220	1	160–180	x	x
Pico116	FTC1UFH02FJEA2.1	TC	32	2	180–200	2	160–180	x	x
Pico138	FTC1UFH02G1YNE	TG	34	2	120–180	3	120–180	x	x
Pico173	FTC1UFH02J0K7H.1	TTTG	36	2	200–220	1	180–200	x	x
Pico174	FTC1UFH02J0M4S.1	TG	40	1	200–220	1	140–160	x	x
Pico183	FTC1UFH02JO6M1.1	TTA	33	1	180–200	1	160–180	x	x
Pico192	FTC1UFH02JZ490.1	TG	32	2	180–200	1	180–200	x	x

Table 3 Number of alleles, minimum (min.) and maximum (max.) allele size, observed (obs.) and expected (exp.) heterozygosity, polymorphism information content (PIC) score, and estimated null allele freq. for 1127 individuals of *Pinus ponderosa*

Locus	No. alleles	Min. size	Max. size	Het obs.	Het exp.	PIC	Est. null allele freq.
Pico2	15	183	211	0.65*	0.78	0.75	0.079
Pico3	11	160	188	0.07*	0.30	0.28	0.290
Pico31	9	162	187	0.07*	0.28	0.27	0.302
Pico104	17	121	169	0.49*	0.89	0.87	0.270
Pico109	25	122	170	0.46*	0.79	0.76	0.174
Pico138	22	126	170	0.43*	0.79	0.78	0.308

*Significant difference between observed and expected heterozygosities at $P < 0.05$.

Estimated null allele frequencies ranged from 0.079 to 0.308 (Table 3). Corrected F_{ST} and Jost's D_{est} values of between population differentiation, based on these six loci, ranged from 0.016 to 0.05, and 0.048 to 0.147, respectively (Table 4). F_{IS} values ranged from 0.401 to 0.542 (Table 4).

Table 4 F_{ST} (above diagonal) and Jost's D_{est} (below diagonal) values across six loci for four populations of *Pinus ponderosa* in north central Wyoming. F_{IS} values are also given for each population

Population	1	2	3	4
1	–	0.027	0.026	0.050
2	0.048	–	0.026	0.018
3	0.079	0.055	–	0.016
4	0.156	0.147	0.063	–
F_{IS}	0.508	0.401	0.441	0.542

Discussion

The development of SSRs in the genus *Pinus* has been focused on a few species. *Pinus taeda* has received the most attention (Chagne *et al.* 2004; Echt *et al.* 2011) and has been used as the basis for other SSR cross-transferability studies in *Pinus* (Chagne *et al.* 2004; Liewlaksaneeyanawin *et al.* 2004). SSR development has occurred in other *Pinus* spp. [see Chagne *et al.* (2004) for review], and in other conifers (Chagne *et al.* 2004; Berube *et al.* 2007), however, these efforts still only represent a small percentage of conifers and the genus *Pinus*. Our results highlight the utility of pyrosequenced EST collections for the identification and characterization of large numbers of gene-based SSR loci across species for which limited marker resources were available (Table 1). The loci detected and tested here represent additional molecular markers to support population genetic, linkage mapping and population genomic studies across the genus *Pinus*.

EST databases have been recognized for their value in the characterization and development of molecular markers (Bouck & Vision 2007; Ellis & Burke 2007), and this should increase with the growing use of next generation sequencing platforms (Hudson 2008; Wheat 2008; Castoe *et al.* 2010). By mining a large collection of pyrosequenced ESTs for *Pinus contorta*, we identified more than 15 000 SSRs and were able to design primers for a large percentage of loci containing SSRs. We selected primers with stringent and similar properties; hence, we were able to produce a set of loci that were likely to amplify readily under standardized PCR conditions. In addition, the putative identity of genes represented by these sequences was known, as many of the SSR containing sequences were BLAST annotated in a previous transcriptome analysis (Parchman *et al.* 2010). Finally, we developed primers for SSRs residing in contig consensus sequences resulting from aligning the *P. contorta* transcriptome sequences to a *P. taeda* NCBI unigene set, offering a potential source of loci with high probability of cross-species transferability.

The percentage of loci that amplified across the different species surveyed here ranged from 73% (*Pinus ponderosa*) to 54% (*Pinus edulis*). As expected, transferability decreased with increasing evolutionary distance from *P. contorta* (Fig. 1). The species with the lowest number of loci amplifying successfully were *Pinus flexilis*, *Pinus longaeva*, and *P. edulis*, which occur in subsection strobus (Gernandt *et al.* 2005; Eckert & Hall 2006).

Low rates of molecular evolution in the genus *Pinus* could facilitate ready cross-transferability of SSRs and other molecular markers (Neale 2007; Neale & Ingvarsson 2008). As EST sequences are typically conserved relative to noncoding DNA, SSRs residing in EST sequences typically benefit from higher amplification rates and higher levels of cross-species transferability (Barbara *et al.* 2007; Ellis & Burke 2007). The high amplification rates across the species tested here indicate substantial cross-species transferability of these and probably other EST-based SSR markers developed from our pyrosequenced EST collection (Table 1). This is consistent with previous studies indicating high cross-species transferability of EST-based SSRs in *Pinus* and other taxa (Eujayl *et al.* 2002; Chagne *et al.* 2004; Liewlaksaneeyanawin *et al.* 2004; Gutierrez *et al.* 2005).

EST-based SSRs are often reported as having lower levels of polymorphism than genomic SSRs (Cho *et al.* 2000; Eujayl *et al.* 2002; Gutierrez *et al.* 2005; Ellis & Burke 2007) However, of 35 EST-based loci tested for polymorphism and allelic variability in *P. contorta* and *P. ponderosa*, 18 and 15, respectively, were polymorphic, and the six loci that we amplified in four populations of *P. ponderosa* showed high levels of variability (Tables 3 and 4). A total of 99 alleles were found across the six loci (average 16.5 alleles per loci), and the PIC scores indicated that the loci were all informative (Botstein *et al.* 1980). This level of variability is in the same range as SSRs from noncoding regions that have previously been used in *Pinus* (Echt *et al.* 1999; Maherali *et al.* 2002) and indicates that EST-based SSRs, in many cases, may be just as polymorphic as non-EST-based loci. Furthermore, F_{IS}, F_{ST} and Jost's D_{est} values (Table 4) were in the same range as differentiation estimates from non-EST-based loci for *P. contorta*, *P. ponderosa* and other conifers across the same spatial range as the populations measured here (Hamrick *et al.* 1992; Latta & Mitton 1999; Parchman *et al.* 2011).

A potential issue in our data was that observed heterozygosity was significantly lower than expected in the assayed loci, and estimated null allele frequencies were high. This is a common issue with SSRs and should be evaulated and addressed where appropriate (Dakin & Avise 2004; Chapuis & Estoup 2007). Null alleles do not have a large effect on linkage mapping applications, but can be an issue for population genetic studies because of underestimation of heterozygosity (van Oosterhout *et al.*

2004). However, corrections can be made to account for null alleles (Chapuis & Estoup 2007), and Carlsson (2008) found that effects of null alleles on assignment tests were minimal. Still, null alleles may bias results, and with the large number of potential candidate SSRs presented here, we suggest caution in the use of loci with these characteristics.

In conclusion, these results complement previous SSR developments in *Pinus* (Chagne *et al.* 2004; Liewlaksaneeyanawin *et al.* 2004; Berube *et al.* 2007; Echt *et al.* 2011) and represent a valuable resource for genetic analysis. We only tested a small subset of the SSR loci identified in a collection of pyrosequenced ESTs available for *P. contorta*, but high amplification rates, and high levels of polymorphism, indicate that the full set of 4000 primer combinations produced for SSRs in Parchman *et al.* (2010) could be a rich source of candidate molecular markers for the genus *Pinus*.

Acknowledgements

This research was funded by an NSF DDIG (DEB-0910173), Global Forest (GLOBLFOR48026) and a USDA McIntire-Stennis Competitive Grant. We thank B. Jenkins for laboratory assistance and R. Jones and C. Edwards for providing collection material.

References

Barbara T, Palma-Silva C, Paggi GM *et al.* (2007) Cross-species transfer of nuclear microsatellite markers: potential and limitations. *Molecular Ecology*, **16**, 3759–3767.

Berube Y, Zhuang J, Rungis D *et al.* (2007) Characterization of EST SSRs in loblolly pine and spruce. *Tree Genetics and Genomes*, **3**, 251–259.

Botstein D, White R, Skolnick M, Davis R (1980) Construction of a genetic-linkage map in man using restriction fragment length ploymorphisms. *American Journal Of Human Genetics*, **32**, 314–331.

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology*, **16**, 907–924.

Carlsson J (2008) Effects of microsatellite null alleles on assignment testing. *Journal of Heredity*, **99**, 616–623.

Castoe TA, Poole AW, Gu W *et al.* (2010) 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.

Chagne D, Chaumeil P, Ramboer A *et al.* (2004) Cross-species transferability and mapping of genomic and cDNA SSRs in pines. *Theoretical and Applied Genetics*, **109**, 1204–1214.

Chapuis M, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621–631.

Cho YG, Ishii T, Temnykh S *et al.* (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics*, **100**, 713–722.

Crawford NG (2010) SMOGD: software for the measurement of genetic diversity. *Molecular Ecology Resources*, **10**, 556–557.

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.

Dieringer D, Schlötterer C (2003) Microsatellite anlalyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, **3**, 167–169.

Doyle J (1991) DNA protocols for plants: a CTAB total DNA isolation. In: *Molecular Techniques in Taxonomy* (eds Hewitt GM, Johnston A), pp. 283–293. Springer-Verlag, Berlin.

Echt CS, Nelson CD (1997) Linkage mapping and genome length in eastern white pine (*Pinus strobus* L.). *Theoretical and Applied Genetics*, **94**, 1031–1037.

Echt CS, Vendramin GG, Nelson CD, Marquardt P (1999) Microsatellite DNA as shared genetic markers among conifer species. *Canadian Journal of Forest Research*, **29**, 365–371.

Echt C, Saha S, Krutovsky K *et al.* (2011) An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. *BMC Genetics*, **12**, 17.

Eckert AJ, Hall BD (2006) Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. *Molecular Phylogenetics and Evolution*, **40**, 166–182.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. *Heredity*, **99**, 125–132.

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics*, **104**, 399–407.

Gernandt DS, López GG, García SO, Liston A (2005) Phylogeny and classification of *Pinus*. *Taxon*, **54**, 29–42.

Guevara MA, Soto A, Collada C *et al.* (2005) Genomics applied to the study of adaptation in pine species. *Investigacion Agraria: Sistemas y Recursos Forestales*, **14**, 292–306.

Gupta PK, Rustgi S, Sharma S *et al.* (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics*, **270**, 315–323.

Gutierrez MV, Patto MCV, Huguet T *et al.* (2005) Cross-species amplification of *Medicago truncatula* microsatellites across three major pulse crops. *Theoretical and Applied Genetics*, **110**, 1210–1217.

Hamrick PL, Godt MJW, Sherman Broyles M (1992) Factors influencing levels of genetic diversity in woody plant species. *New Forests*, **6**, 95–124.

Hempel K, Peakall R (2003) Cross-species amplification from crop soybean *Glycine max* provides informative microsatellite markers for the study of inbreeding wild relatives. *Genome*, **46**, 382–393.

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

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Kim KS, Ratcliffe ST, French BW, Liu L, Sappington TW (2008) Utility of EST-derived SSRs as population genetics markers in a beetle. *Journal of Heredity*, **99**, 112–124.

Latta RG, Mitton JB (1999) Historic separation and present gene flow through a zone of secondary contact in ponderosa pine. *Evolution*, **53**, 769–776.

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. *Theoretical and Applied Genetics*, **109**, 361–369.

Maherali H, Williams B, Paige K, Delucia E (2002) Hydraulic differentiation of Ponderosa pine populations along a climate gradient is not associated with ecotypic divergence. *Functional Ecology*, **16**, 510–521.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.

Moccia MD, Oger-Desfeux C, Marais GAB, Widmer A (2009) A White Campion (*Silene latifolia*) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping. *BMC Genomics*, **10**, 1–14.

Neale DB (2007) Genomics to tree breeding and forest health. *Current Opinion in Genetics and Development*, **17**, 539–544.

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology*, **11**, 149–155.

Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Ong C, Yusoff K, Yap C, Tan S (2009) Genetic characterization of *Perna viridis* L. in peninsular Malaysia using microsatellite markers. *Journal of Genetics*, **88**, 153–163.

van Oosterhout C, Hutchinson W, Willis D, Shipley P (2004) Microchecker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.

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.

Parchman TL, Benkman CW, Jenkins B, Buerkle CA (2011) Low levels of population genetic structure in *Pinus contorta* (Pinaceae) across a geographic mosaic of co-evolution. *American Journal of Botany*, **98**, 669–679.

Ralph SG, Yueh H, Friedmann M *et al.* (2006) Conifer defence against insects: microarray gene expression profiling of sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. *Plant Cell and Environment*, **29**, 1545–1570.

Schlötterer C (1998) Genome evolution: are microsatellites really simple sequences? *Current Biology*, **8**, R132–R134.

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

Simko I (2009) Development of EST-SSR markers for the study of population structure in Lettuce (*Lactuca sativa* L.). *Journal of Heredity*, **100**, 256–262.

Slate J, Hale MC, Birkhead TR (2007) Simple sequence repeats in zebra finch (*Taeniopygia guttata*) expressed sequence tags: a new resource for evolutionary genetic studies of passerines. *BMC Genomics*, **8**, 1–12.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.

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

Vasemagi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology and Evolution*, **22**, 1067–1076.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution*, **38**, 1358–1370.

Wheat CW (2008) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*, **138**, 433–451.

Woodhead M, Russell J, Squirrell J *et al.* (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. *Molecular Ecology*, **14**, 1681–1695.

You FM, Huo N, Gu YQ *et al.* (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, **9**, 253.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology*, **11**, 1–16.

Data accessibility

Files containing the sequences and quality scores for the raw 454 pyrosequencing reads have been deposited at NCBIs Short Read Archive (accession SRA012089). Consensus sequences of contigs and singleton sequences used for SSR detection and characterization are deposited at DRYAD (doi: 10.5061/dryad.8dk6t78s).

Supporting information

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

Appendix S1 Primer sequences, annealing temperatures, product size, repeat motif, number of repeats and expected product lengths for 4020 SSRs detected in *P. contorta* pyrosequenced ESTs taken from Parchman *et al*. (2010).

Appendix S2 Primer sequences, annealing temperatures, product size, repeat motif, number of repeats and expected product lengths for the 184 loci tested in this study, and whether that locus amplified in each of the nine species tested.

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.