Tree Genetics & Genomes (2011) 7:773–780
DOI 10.1007/s11295-011-0373-7

ORIGINAL PAPER

# Microsatellite DNA in genomic survey sequences and UniGenes of loblolly pine

**Craig S. Echt · Surya Saha · Dennis L. Deemer · C. Dana Nelson**

Received: 23 September 2010 / Revised: 4 January 2011 / Accepted: 7 February 2011 / Published online: 1 March 2011


**Abstract** Genomic DNA sequence databases are a potential and growing resource for simple sequence repeat (SSR) marker development in loblolly pine (*Pinus taeda* L.). Loblolly pine also has many expressed sequence tags (ESTs) available for microsatellite (SSR) marker development. We compared loblolly pine SSR densities in genome survey sequences (GSSs) to those in non-redundant EST and cDNA sequences (UniGenes) and designed SSR primer pairs from both sequence types. Overall SSR densities were 96 SSR/Mb in GSSs and 38 SSR/Mb in UniGenes. Loblolly pine had the lowest transcriptome SSR density when compared to 49 other species in the NCBI UniGene database. Among the five different GSS genome fractions of loblolly pine analyzed, methylation-filtered DNA had the highest SSR density at 145 SSR/Mb. The most abundant loblolly pine SSR motif was AT in both GSSs (34 SSR/Mb) and UniGenes (7.6 SSR/Mb). Among the trinucleotide SSR motifs, the most abundant was AAT (9.5 SSR/Mb) in GSSs and AGC (4.9 SSR/Mb) in UniGenes. We designed PCR primer pairs for 120 genomic SSRs and 315 EST-SSRs and evaluated PCR amplification for 108 (25%). We identified 21 primer pairs that reliably amplified polymorphic loci from 31 loblolly pine individuals and estimated that at least 60 additional polymorphic marker loci could be developed from available *P. taeda* GSS and UniGene resources.

**Keywords** EST-SSR · GSS-SSR · Microsatellite marker · *Pinus taeda* · SSR frequency

Communicated by S. González-Martínez

**Electronic supplementary material** The online version of this article (doi:10.1007/s11295-011-0373-7) contains supplementary material, which is available to authorized users.

C. S. Echt (✉) · D. L. Deemer · C. D. Nelson
Southern Institute of Forest Genetics,
Southern Research Station U.S. Forest Service,
Saucier, MS 39574, USA
e-mail: cecht@fs.fed.us

S. Saha
Mississippi Genome Exploration Laboratory,
Mississippi State University,
Mississippi State, MS 39762, USA

S. Saha
Department of Computer Science and Engineering,
Mississippi State University,
Mississippi State, MS 39762, USA

S. Saha
Department of Plant Pathology and Plant-Microbe Biology,
Cornell University,
Ithaca, NY 14853, USA

**Abbreviations**

| | |
|---|---|
| EST | Expressed sequence tag |
| GSS | Genome survey sequence |
| NCBI | National Center for Biotechnology Information |
| Mb | Megabase |
| SSR | Simple sequence repeat |

## Introduction

Microsatellite sequences, also known as simple sequence repeats (SSRs), are interesting components of genomes because they function in gene regulation (Zhang et al. 2006), recombination (Bagshaw et al. 2008), and evolvability (Vinces et al. 2009), as well as provide a rich source of informative genetic markers for a variety of applications (Ellegren 2004; Ellis and Burke 2007). A traditional approach for SSR marker development has

been to select and individually sequence SSR-bearing clones from genomic libraries; however, for the large and complex genomes found in pine (*Pinus*) and spruce (*Picea*) species, obtaining markers by this approach is relatively inefficient (A'Hara and Cottrell 2009; Echt and May-Marquardt 1997; Elsik and Williams 2001; Fisher et al. 1996; Pfeiffer et al. 1997; Soranzo et al. 1998; Van de Ven and McNicol 1996; Zhou et al. 2002). Consequently, alternate sources of SSR sequences should be considered when there is a need to develop SSR markers from such species. Loblolly pine (*Pinus taeda* L.) is an economically and ecologically important tree species in the southern USA for which we are developing SSR markers beyond what are currently available (Auckland et al. 2002; Chagné et al. 2004; Liewlaksaneeyanawin et al. 2004).

Potential alternate sources of DNA sequences for SSR marker development include public genomic DNA and EST databases and de novo high-throughput next-generation sequencing of genomic DNA and EST (cDNA) libraries. Genomic DNA sequences have proven to be an important source of SSR markers in *Populus* (Yin et al. 2009), and genomic shotgun sequences have recently become available for *P. taeda* through the genome survey sequence (GSS) database (NCBI 2010). Transcriptome sequences in the form of ESTs are another important source of SSR markers for a number of plant species (Varshney et al. 2005). Although conifer ESTs have comparatively few SSRs (Bérubé et al. 2007; Chagné et al. 2004; Liewlaksaneeyanawin et al. 2004; von Stackelberg et al. 2006), over 328,000 *P. taeda* ESTs and mRNA sequences, grouped into about 18,000 non-redundant UniGene clusters, are available (NCBI 2010). If DNA database resources are insufficient or lacking for a particular species, then de novo shotgun genomic, SSR-enriched genomic, or transcriptomic sequencing can rapidly and cost-effectively provide SSR sequence data needed for SSR marker development (Abdelkrim et al. 2009; Csencsics et al. 2010; Martin et al. 2010; Mikheyev et al. 2010; Parchmann et al. 2010; Tangphatsornruang et al. 2009). Finally, SSR markers also can be obtained opportunistically by evaluating PCR primer pairs developed from related species (Barbará et al. 2007; Ellis and Burke 2007). For pine genomic SSR markers, reported transfer rates range from 57% to 100% within taxonomic subsections and 25% to 86% between subsections (Chagné et al. 2004; Echt et al. 1999; Devey et al. 1999; González-Martínez et al. 2004; Kutil and Williams 2001; Liewlaksaneeyanawin et al. 2004; Shepherd et al. 2002). For *P. taeda* EST-SSR markers, reported transfer rates to various pine species range from 65% to 100% (Chagné et al. 2004; Liewlaksaneeyanawin et al. 2004; Shepherd et al. 2002). Most SSR marker development for pines has been with *P. taeda*, which limits options to transfer markers from other pine species to *P. taeda*.

Whether mining databases or generating de novo sequences, SSR marker development efforts would proceed most rapidly from sources that have the greatest abundance of SSRs. While EST-SSR frequency has been measured for the *Pinus* Gene Index (von Stackelberg et al. 2006), the only reported measurement of SSR frequency in pine genomic DNA has been with plaque lift hybridizations of SSR probes to a *P. taeda* genomic phage library (Echt and May-Marquardt 1997). To provide a direct comparison of SSR frequencies among different sequenced genome fractions, we measured SSR frequencies in *P. taeda* GSS libraries, *P. taeda* UniGenes, and UniGenes from other plants. To provide useable SSR markers, we designed PCR primer pairs from *P. taeda* GSS and UniGene sequences and evaluated a subset of the primers for SSR marker amplification and polymorphism in *P. taeda*.

## Materials and methods

### DNA sequences

*P. taeda* genomic sequences were downloaded from the GSS database division of GenBank (http://www.ncbi.nlm.nih.gov/dbGSS/) and are referred to here as GSSs. All sequences were provided by two laboratories, The Institute for Genomic Research (TIGR) and Mississippi Genome Exploration Laboratory (MGEL), using standard Sanger capillary sequencing methodology. TIGR sequences were from methylation-unfiltered and methylation-filtered genomic libraries, which target random and gene-rich DNA fragments, respectively (Rabinowicz et al. 2005). MGEL sequences were from unfractionated, high-$C_o$t-fractionated, and mid-$C_o$t-fractionated genomic libraries, which target random, single/low-copy, and moderately repetitive DNA fragments, respectively (Peterson et al. 2002).

Non-redundant *P. taeda* UniGene sequences were obtained from the UniGene database division of NCBI as a Pta.seq.uniq.gz file from ftp://ftp.ncbi.nih.gov/repository/UniGene/Pinus_taeda/. Each unique sequence in this file represented a UniGene cluster of sequences as the EST with the longest high-quality sequence or as the annotated mRNA sequence. For ease of discussion, we here refer to those sequences simply as UniGenes. UniGenes from all other available plant species in the database were also obtained from ftp://ftp.ncbi.nih.gov/repository/UniGene/.

### SSR selection and statistical analysis

Sequences were analyzed by the software SciRoKo (Kofler et al. 2007) using the perfect-MISA search mode with program parameters set to 11-6-5-4-4-4 (indicating the respective minimum number of repeats for mononucleotide

through to hexanucleotide motifs). SciRoKo reporting parameters were set as follows: upper motif length=6, lower motif length=2, minimum required score (SSR length)=12, and minimum count=1. We also counted perfect SSRs that met the same motif and length criteria with the PERL script for SSRIT (Temnykh et al. 2001, http://www.gramene.org/db/markers/ssrtool) and verified that the two analysis procedures yielded equivalent SSR counts. Nomenclature for SSR motifs followed the convention of alphabetical simplification (Echt and May-Marquardt 1997), which is the same as the "fully standardized" nomenclature used by SciRoKo (Kofler et al. 2007). For example, all SSRs of the type $(AAT)_n$, $(ATA)_n$, and $(TAA)_n$ and their reverse complements $(ATT)_n$, $(TAT)_n$, and $(TTA)_n$ are referred to collectively as an $(AAT)_n$ motif or, in simpler form, as an AAT motif.

We standardized SSR abundance as SSR density, which is the number of SSRs per Mb of analyzed sequence. To statistically test for measured differences in SSR densities between DNA sources or species, we conducted Chi-square ($\chi^2$) analyses of differences in SSR frequencies between reference and contrast sets of DNA sequences. We did this by counting for each set the number of "sequence units", $S$, with and without an SSR. We defined the $S_{tot}$ total number of sequence units in a set as, $S_{tot}=L_{tot}/L_{ssr}$, where $L_{tot}$ was the total number of nucleotides in the set and $L_{ssr}$ was a specified SSR nucleotide length. For $L_{ssr}$, we used the empirically determined average SSR length: 19.24 for all sets of plant UniGenes and 23.24 for all sets of *P. taeda* GSSs, though any $L_{ssr}$ value in that general range would have provided comparable test results. A set's $S_{ssr}$ number of units with an SSR was its SSR count, while its $S_{non}$ number of units without an SSR was calculated as $S_{non}=S_{tot}-S_{ssr}$. We used the proportion of SSRs in a reference set, $S_{ssr}/S_{tot}$, to estimate the expected values of $S_{ssr}$ and $S_{non}$ for a contrast set, given the contrast set's $S_{tot}$. The expected and observed values of $S_{ssr}$ and $S_{non}$ for a contrast set were used to test, under the null hypothesis and with one degree of freedom, whether it had the same SSR frequency as the reference set. We calculated $\chi^2$ statistics and $P$ values using spreadsheet functions and using an on-line calculator (Kirkman 1996).

## PCR primer design and selection

We specified SSR targets for PCR primer design using the same criteria as for counting SSRs described above. Primer pairs were designed from non-redundant sequences with the STS_Pipeline1.3 package, which we modified from STS_Pipeline1.2 (Resnick and Stein 1995) by porting it to the Redhat Linux 9.0 operating system and updating its primer design module to Primer3 (http://primer3.sourceforge.net/). Input parameters used for the primer design were as follows: minimum size=18 nt, maximum size=24 nt, optimal size=20 nt, minimum GC content=20%, maximum GC content=80%, minimum $T_m$=56°C, maximum $T_m$=64°C, optimal $T_m$=60°C, maximum 3′ end complementarities=3 nt, and maximum any complementarities=8 nt.

Three primer pairs, those having the top ranking PRIMER_PAIR_QUALITY output tags (quality scores), were reported for each sequence. We screened results with custom PERL scripts to select one primer pair from each sequence that flanked the longest target SSR and had an amplicon size and quality score that were within our target marker size ranges. For example, for a dinucleotide SSR, the primer pair with an amplicon size between 100 and 300 bp was selected if that pair had the best quality score of the three alternatives; otherwise, the pair with the smallest amplicon size between 300 and 500 bp was selected. For longer SSR motifs, the preference was for an amplicon size between 300 and 500 bp if the primer pair had the best quality score; otherwise, the longest amplicon of the three alternatives was selected. We used this size-distributed selection strategy to increase the number of potential marker loci that we could use in pooled genotyping assays on capillary electrophoresis platforms.

## SSR marker evaluation

DNA was isolated from 50 mg finely chopped frozen needle tissue using DNeasy 96 Plant Kits (QIAGEN, Inc) and quantified by Hoescht dye fluorometry. All marker-specific PCR forward primer oligonucleotides were synthesized with the M13 forward (−29) sequence (CACGACGTTGTAAAACGAC) on their 5′ end and all reverse primer oligonucleotides were synthesized with GTTTCTT on their 5′ end (Browenstein et al. 1996). A fluorescent dye (6-FAM, VIC, NED or PET) was incorporated into amplicons by including a 5′ dye-labeled universal M13 forward (−29) primer in each reaction (Schuelke 2000). PCR was performed in a 6-µl volume of the following composition: 10 ng template DNA, 80 nM forward primer, 320 nM reverse primer, 320 nM dye-labeled M13 primer, 66 µM dNTPs, 1.5 U Klear Taq polymerase (KBioscience), buffer "B" (KBioscience), and 1.8 mM $MgCl_2$. We used a hot-start PCR thermocycling protocol, on a Bio-Rad PTC-200, as follows: 94°C (2 min); 20 cycles of 94°C (30 s), 65°C minus 0.5°C per cycle (30 s), 72°C (1 min); 25 cycles of 92°C (30 s), 55°C (30 s), 72°C (1 min 30 s); and 72°C (15 min). Reactions were kept at 4°C until analyzed. We diluted completed reactions 30-fold and pooled four markers by expected amplicon sizes, one for each dye label, to minimize overlap of allele size ranges. One

microliter of pooled product, mixed with ABI PRISM GS 600 LIZ internal size standards (Life Technologies), was separated by capillary electrophoresis using an ABI PRISM 3130xl Genetic Analyzer (Life Technologies) fitted with a 36-cm 96-capillary array and ABI PRISM POP-7 polymer (Life Technologies). We used the default ABI PRISM Foundation Data Collection v.3.1 run module parameters for the given capillary length and polymer type. Amplicon size determination by the third order least-squares algorithm and allele size binning were performed with ABI PRISM GENEMAPPER 3.7 software (Life Technologies). We inspected and edited allele calls as needed and standardized SSR marker allele bins among capillary electrophoresis runs with the aid of *P. taeda* reference samples that were included in each capillary run (Deemer and Nelson 2010).

We evaluated PCR amplification for 25% of SSR primers pairs designed from each sequence class (GSSs and UniGenes) by using samples of our two reference genotypes, 7-56 (coastal South Carolina origin) and B-145-L (southeast Texas origin), and four widely used *P. taeda* reference mapping parents, 11-1060, 20-1010, 6-1031, and 8-1070. Primers that successfully amplified potential marker loci were further evaluated for marker polymorphism among an additional 25 *P. taeda* genotypes that we sampled from across the range of the species. Estimates of null allele frequencies and null-adjusted SSR allele frequencies were obtained by a maximum likelihood method using the EM algorithm (Dempster et al. 1977), as implemented in Option 8.1 of the software program GENEPOP 4.0.7 (Rousset 2007). We modified the standard GENEPOP input data file to include a dummy population of a single sample with all locus genotypes coded as 9999, which is the explicit GENEPOP code for null homozygotes. This modification suppressed the program's default interpretation of the largest allele integer as a null allele and ensured correct allele counts and frequency estimates. We used null-adjusted estimated allele frequencies to calculate for each locus the observed frequency of heterozygosity, $H_o$, and the unbiased expected heterozygosity, $H_e$ (Nei 1978). The arithmetic mean of $H_e$ was used as its multilocus expectation.

## Results and discussion

We analyzed 5,393 *P. taeda* GSSs covering 2.5 Mb and found 242 SSRs in 194 (3.6%) of the sequences. We observed that SSR densities, measured as SSR/Mb, differed among GSS libraries constructed from different genome fractions and that the largest difference was between the methylation-filtered and high-$C_o$t libraries (Table 1). This difference between the two libraries held for the dimer through pentamer motif classes: neither library, nor the mid-$C_o$t library, contained hexamer motifs (Fig. 1). For all GSS genome fractions, dinucleotide SSRs comprised the most abundant SSR motif class. Among the dinucleotide SSRs, AT (34 SSR/Mb) was the most abundant motif followed by AG (20 SSR/Mb), while among the trinucleotide SSRs, AAT (9.5 SSR/Mb) was the most abundant followed by AAG (4.8 SSR/Mb). The most abundant tetranucleotide SSR motif was AAAT (3.6 SSR/Mb). Summary results and detailed data for all GSS SSRs are in Supplementary Table S1.

Repetitive DNA in plants tends to be highly methylated; therefore, removal of these sequences through methylation filtration can enrich for hypomethylated, low-copy, genic DNA (Rabinowicz et al. 2005). Alternatively, low-copy and single-copy genic DNA can be enriched by virtue of its high-$C_o$t values (Peterson et al. 2002; Lamoureux et al. 2005). While the pine methylation-filtered and high-$C_o$t libraries were theoretically enriched for low-copy/single genic sequences, they had not been evaluated to quantify the degree of enrichment. Sequences from the two random genomic library types (TIGR methylation-unfiltered and MGEL genomic DNA) had different SSR densities (Table 1), though the difference was not significant ($\chi^2$=0.20, $P$=0.65, MGEL sequences used as reference set). We deemed these two libraries as equivalent and calculated an SSR density of 92 SSR/Mb for the combined random genomic fractions. The relatively greater abundance of SSRs that we observed in methylation-filtered genome fractions supports the conclusion of a prior report that SSRs are preferentially associated with low-copy (non-repetitive) DNA in plant genomes (Morgante et al. 2002). It also supports observations that *P. taeda* low-copy genomic libraries yield more polymorphic SSR markers than do random genomic (control) libraries (Elsik and Williams 2001). We cannot explain, however, the relatively poorer representation of SSRs in the high-$C_o$t fraction GSSs that, in theory, should have represented a similar genic fraction as methylation-filtered DNA.

To compare GSS-SSR densities with those in transcribed sequences, we analyzed 18,079 *P. taeda* UniGenes covering 14.4 Mb and found 545 SSRs in 500 (2.8%) of the UniGenes. The average SSR density among UniGenes was 38 SSR/Mb, or 39% that found among random GSSs (Table 1). The trinucleotide SSR motif class was the most abundant among UniGenes (Fig. 1). The trinucleotide SSR density in UniGenes (19 SSR/Mb) was greater than it was in total GSSs, though it was less than the density in either the methylation filtration or mid-$C_o$t GSS fractions (Fig. 1). As with GSS SSRs, the most abundant UniGene SSR was AT (7.6 SSR/Mb). The most abundant UniGene trinucleotide SSR motif was AGC (4.9 SSR/Mb), while AAAT (1.0 SSR/Mb) was the most abundant tetranucleotide motif. The second most abundant UniGene trinucleotide motif was

**Table 1** Comparison of SSR abundance among *P. taeda* genome fractions

| Genome fraction, database | No. of sequences analyzed | Avg. sequence length, bp | Total Mb | No. of SSRs[a] | SSR/Mb |
|---|---|---|---|---|---|
| Methylation filtered, GSS | 614 | 630 | 0.387 | 56 | 145 |
| High-$C_o$t, GSS | 2,328 | 213 | 0.496 | 32 | 65 |
| Mid-$C_o$t, GSS | 266 | 442 | 0.117 | 14 | 119 |
| Random (TIGR), GSS | 1,178 | 546 | 0.643 | 57 | 89 |
| Random (MGEL), GSS | 1,007 | 876 | 0.882 | 83 | 94 |
| All, GSS | 5,393 | 468 | 2.522 | 242 | 96 |
| cDNA, UniGene | 18,079 | 798 | 14.424 | 545 | 38 |

[a] Perfect SSRs counted: $(NN)_{\geq 6}$, $(NNN)_{\geq 5}$, $(NNNN)_{\geq 4}$, $(NNNNN)_{\geq 4}$, $(NNNNNN)_{\geq 4}$

AAG (4.2 SSR/Mb), which was also the second most abundant trinucleotide motif among GSS SSRs. Summary results and detailed data for all UniGene SSRs are in Supplementary Table S2.

We also compared the UniGene SSR density of *P. taeda* to that of 49 other plant species in the NCBI UniGene database (Supplementary Table S3). *P. taeda* had the lowest average SSR density among the 50 species. Based on the current UniGene sampling, *P. taeda* SSR density did not statistically differ from that of the next lowest ranked species, *Festuca pratensis* (43 SSR/Mb, $\chi^2 = 2.3$, $P = 0.13$, *P. taeda* used as reference set), though it did differ from the other two other conifers analyzed, *Picea glauca* (47 SSR/Mb, $\chi^2 = 39.5$, $P = 3.3 \times 10^{-10}$) and *Picea sitchensis* (48 SSR/Mb, $\chi^2 = 51.5$, $P = 7.0 \times 10^{-13}$), which were third and fourth lowest ranked among the 50 species. The SSR densities of these two spruce species, however, did not significantly differ from each other ($\chi^2 = 0.55$, $P = 0.46$, *P. sitchensis* used as reference set).

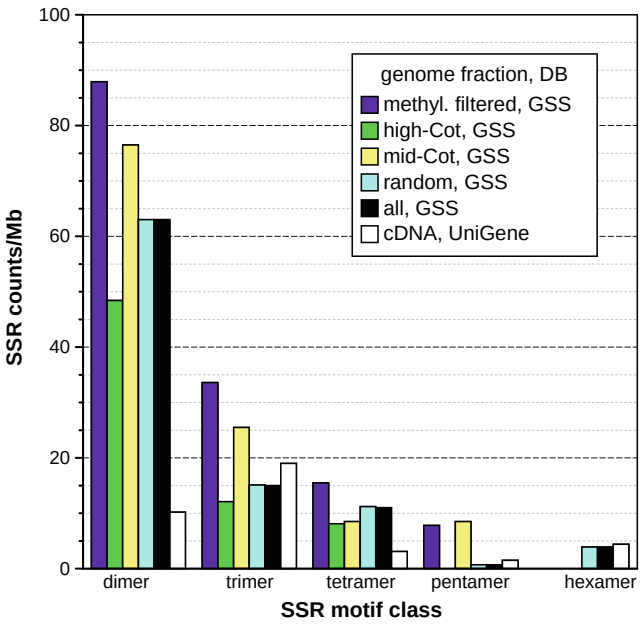


**Fig. 1** *Bar graph* of SSR densities (SSR/Mb), by five SSR motif classes for each genome fraction in *P. taeda* GSS and UniGene sequences

A previous plant SSR density survey used EST contig sequences from 24 plant taxa, most of which were from the Plant Gene Index (von Stackelberg et al. 2006). That study ranked *Pinus* as the third lowest taxon with 41 SSR/Mb—only *Mesostigma* (37 SSR/Mb) and *Allium* (39 SSR/Mb) ranked lower—and *Picea* ranked fourth lowest with 47 SSR/Mb. Nearly all *Pinus* ESTs in public databases at the time of that study were from *P. taeda*; therefore, we interpret the prior reported SSR density for the *Pinus* Gene Index as indicative of *P. taeda*. Our measured *P. taeda* density of 38 SSR/Mb differed slightly from that prior report most likely because of the different SSR enumeration parameters used and different sets of sequences analyzed, though the rankings of those taxa held in common between both surveys are generally comparable. Based on current data, we conclude that the transcribed genome of *P. taeda* has among the lowest SSR densities for plant species.

We designed PCR primer pairs for SSR marker loci from 120 (2.2%) of the 5,393 GSSs and 315 (1.7%) of the 18,079 UniGenes (Supplementary Table S4). None of our GSS-SSR primer pairs targeted previously reported *P. taeda* genomic SSR marker loci (Chagné et al. 2004; Echt et al. 2011; Auckland et al. 2002), and only five of our UniGene EST-SSR primer pairs targeted sequences that were homologous to previously reported *P. taeda* EST-SSR loci (Chagné et al. 2004; Echt et al. 2011; Liewlaksaneeyanawin et al. 2004), as noted in Supplementary Table S4. Therefore, 430 of our new primer pairs were novel with respect to previous *P. taeda* SSR marker loci. We found that 47% (14 out of 30) of evaluated GSS-SSR primer pairs provided what appeared to be single-locus amplification in the expected size ranges, as did 54% (42 out of 78) of evaluated EST-SSR primer pairs. In comparison, Zhou et al. (2002) reported a statistically equivalent amplification success rate of 46% (113 out of 245) from hypomethylated genomic DNA libraries ($\chi^2 = 0.029$, $P = 0.86$). Guevara et al. (2005) reported an amplification success rate of 54% (24 out of 58) from *Pinus pinaster* genomic SSR primers, which also does not

**Table 2** Characteristics of 21 polymorphic SSR loci, including assigned marker name, GenBank sequence accession number, the targeted SSR in the GenBank sequence, PCR primer pair sequences, UniGene cluster identifier for the GenBank sequence, range of observed marker sizes among 31 *P. taeda* individuals, number of visible SSR alleles (A), estimated null allele frequency (Null), unbiased null-adjusted observed heterozygosity ($H_o$), and unbiased null-adjusted expected heterozygosity ($H_e$)

| Marker name | GenBank acc. no. | SSR | PCR primer sequences (5′–3′) | UniGene ID | Marker size range, bp | A | Null | $H_o$ | $H_e$ |
|---|---|---|---|---|---|---|---|---|---|
| PtSIFG_5015 | CZ894980 | $(AAT)_8$ | F: AAGTCAAAGGTCAAATCATTATGC<br>R: AAACAAGCCACACAAACACG | n/a | 388–405 | 7 | 0.116 | 0.832 | 0.814 |
| PtSIFG_5020 | CZ895059 | $(AG)_{21}$ | F: CACCGCGTTGATGTATTGAA<br>R: TCCATGGTTTTCTCCTCTGAA | n/a | 420–507 | 19 | 0.320 | 0.967 | 0.883 |
| PtSIFG_6004 | AW290058 | $(TA)_6$ | F: AGGCAGTCTTGTCCTCTGGA<br>R: CTATTTTCTGCAAGCTGGGC | Pta.3251 | 290–291 | 2 | 0.000 | 0.542 | 0.472 |
| PtSIFG_6013 | BQ699280 | $(AGC)_5$, $(AGG)_5$ | F: TCCTTATGTTTTCAAGGCCG<br>R: CTGGGAAATTGTGACCTCGT | Pta.12837 | 414–424 | 6 | 0.028 | 0.613 | 0.502 |
| PtSIFG_6014 | CF386356 | $(GAG)_5$ | F: GGAAGCGATTCCCACAACTA<br>R: CTTTCGGAGGTTCAGAGGTG | Pta.5677 | 438–444 | 3 | 0.000 | 0.066 | 0.064 |
| PtSIFG_6015 | CF386735 | $(TA)_7$, $(AT)_6$ | F: TGCTCCAACTGTCTCCACAC<br>R: GCCTAGGCCGAAGATAAACC | Pta.15081 | 321–368 | 18 | 0.125 | 0.910 | 0.930 |
| PtSIFG_6021 | CF396247 | $(GAACCT)_4$ | F: TCACCATCATCCCTCTCCTC<br>R: GAGCCAAACGGTTTTCATGT | Pta.2364 | 399–411 | 3 | 0.000 | 0.230 | 0.210 |
| PtSIFG_6028 | CF668467 | $(AT)_7$ | F: TGCAGTCAATGTGCATGAGA<br>R: TTAACTCCTGGAGGTGGTGG | Pta.17303 | 256–258 | 2 | 0.000 | 0.034 | 0.033 |
| PtSIFG_6030 | CK604170 | $(AT)_{19}$ | F: GGTTTGATGGGAAAGCTTCA<br>R: ATGCTCATCCGAAGACTGCT | Pta.12945 | 375–406 | 15 | 0.238 | 0.925 | 0.894 |
| PtSIFG_6034 | CN785731 | $(TTC)_5$ | F: CTCGATTTTCACCGCTCTTC<br>R: AGGCGCAGTATCCAGAAAGA | Pta.960 | 478–484 | 2 | 0.000 | 0.066 | 0.063 |
| PtSIFG_6035 | CO157652 | $(GAA)_6$ | F: CGCAAGGAGGAGATGTTGAC<br>R: GGGGACAAATAAGCTAGCCC | Pta.14903 | 442–451 | 4 | 0.014 | 0.344 | 0.370 |
| PtSIFG_6039 | CO161878 | $(TAAA)_4$ | F: AACATCTGCCGACTACCCAC<br>R: GTTTTATCCCGATCATTGCG | Pta.15547 | 382–394 | 2 | 0.000 | 0.491 | 0.460 |
| PtSIFG_6044 | CO164176 | $(TA)_8$ | F: TGATGGGTGCCAGTAACAAA<br>R: AATTGACCAATCGTGCGTCT | Pta.7672 | 235–248 | 7 | 0.000 | 0.610 | 0.564 |
| PtSIFG_6046 | CO164853 | $(ATTT)_4$ | F: CTGAAAACCCATTCTTCCCA<br>R: ACGGAGTACCTCATCAACGG | Pta.17984 | 313–334 | 5 | 0.093 | 0.699 | 0.715 |
| PtSIFG_6050 | CO167067 | $(CTT)_5$ | F: GCCATTTCTTCTGCTTCCAG<br>R: AAGGGCTCTTTTCACCCATT | Pta.9088 | 210–213 | 2 | 0.068 | 0.273 | 0.283 |
| PtSIFG_6055 | CO199761 | $(TA)_{12}$ | F: TGTTGGCTACAATTCCCACA<br>R: AAATCATCCCAAACACGAGC | Pta.18276 | 253–269 | 9 | 0.333 | 0.906 | 0.833 |
| PtSIFG_6056 | CO199953 | $(CATT)_4$ | F: TCAAATTCCAACAATGGGGT<br>R: AGCAGGAGATCAGCCAATGT | Pta.3499 | 345–357 | 2 | 0.000 | 0.175 | 0.160 |
| PtSIFG_6058 | CO361898 | $(AT)_6$, $(GAC)_5$ | F: AAGAGGTTGCTCCTCACCAA<br>R: GCTCCATTTCAGAGCAGGTC | Pta.18785 | 144–165 | 5 | 0.069 | 0.366 | 0.360 |
| PtSIFG_6059 | CO363332 | $(TA)_7$ | F: GGCGAAATAGATAAAACCATGC<br>R: CAGTTTGCGAGGAGACAACA | Pta.17188 | 221–253 | 11 | 0.222 | 0.853 | 0.839 |
| PtSIFG_6065 | CO365046 | $(AT)_8$ | F: TATTCCGGTAGTCTGGCACC<br>R: GAAGGAAGCAGCTTTTGCAC | Pta.17251 | 204–222 | 10 | 0.000 | 0.918 | 0.851 |
| PtSIFG_6072 | CO368277 | $(GTC)_5$ | F: GAAGAAGCTGCTGAATTGGG<br>R: ACAACTGCATTAAGCACCCC | Pta.2401 | 445–453 | 3 | 0.000 | 0.066 | 0.064 |

significantly differ from our mean value of 47% ($\chi^2$=0.736, $P$=0.39). For EST-SSR primers pairs, Liewlaksaneeyanawin et al. (2004) reported an amplification success rate of 50% (7 out of 14) from *P. taeda*, which does not differ from our mean of 54% ($\chi^2$=0.123, $P$=0.73). In contrast, Chagné et al. (2004) reported a much higher rate of 75% (52 out of 69) from *P. taeda* and Parchman et al. (2010) reported a rate of 70% (67 out of 96) from *Pinus contorta*. A recent genome-wide analysis of SSR markers from *Populus tremuloides* found, among 150 marker loci tested, an amplification success rate of 63% for primers in intronic and intergenic regions and a rate of 80% for primers in exonic regions (Yin et al. 2009). Together these results suggest that from new SSR primers, higher amplification success rates can be expected from EST-SSRs than from genomic SSRs. They also suggest that higher amplification rates than what we found are possible. We propose that further optimization of our PCR conditions could recover additional SSR markers from the set of primer pairs that we evaluated.

In a population of 31 *P. taeda* individuals sampled from across the geographic range, we obtained clear amplification of two or more alleles for 2 of the 30 GSS-SSR loci and 19 of the 78 EST-SSR loci (Table 2). Among the 21 polymorphic marker loci, the average number of alleles was 6.5 and the average gene diversity, $H_e$, adjusted for null alleles, was 0.49. The presence of null alleles was estimated, though not verified by segregation tests, for 13 loci (Table 2). To provide a calibration reference for subsequent studies using these loci, SSR allele genotypes for our six evaluation samples are given in Supplementary Table S5. We also found six monomorphic loci (noted in Supplementary Table S4) that may prove useful as species-specific markers for *P. taeda* if upon subsequent evaluations they are found to be fixed for different alleles in other pine species. Based on our observed rates of recovery of polymorphic markers, we estimate that at least 60 additional polymorphic SSR marker loci can be developed from our remaining unevaluated 325 primer pairs.

Twenty-nine of the 30 GSS-SSR primer pairs that we evaluated were from methylation-filtered (hypomethylated) sequences, providing a rate of recovery of approximately 0.3% for polymorphic markers from sequences of that low-copy genomic fraction. We did not evaluate SSR polymorphism for markers from the random GSS fractions, though Elsik and Williams (2001) did report a polymorphic marker recovery rate from random genomic DNA that was about half that from hypomethylated genomic DNA. Assuming that reported difference is applicable to the GSS genomic fractions that we analyzed, we could expect a rate of about 0.15% from random GSSs. Our observed rate of polymorphic marker recovery from UniGenes was lower at 0.1%. These and prior results suggest that polymorphic *P. taeda* SSR markers could be developed more efficiently from genomic (preferably hypomethylated) DNA sequences than from transcribed sequences. While this may appear to contradict reports that marker polymorphism is lower among pine EST-SSRs than it is among genomic SSRs (Chagné et al. 2004; Liewlaksaneeyanawin et al. 2004), there is no inherent discrepancy because prior studies measured polymorphism rates from the number of evaluated available SSR primer pairs, not from the number of originating EST and genomic sequences. Alternatively, if the goal were to develop *P. taeda* SSR markers for use in other pine species, then the higher interspecies transferability of EST-SSRs would provide greater marker development efficiency from transcribed sequences than from genomic sequences (Chagné et al. 2004; Ellis and Burke 2007; Liewlaksaneeyanawin et al. 2004).

**Acknowledgments** We thank Jim Roberds for his helpful discussions during the preparation of this manuscript and thank Tom Kubisiak, Ross Whetten, and an anonymous reviewer for their insightful comments on previous versions.

## References

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46:185–192

A'Hara SW, Cottrell JE (2009) Development of a set of highly polymorphic genomic microsatellites (gSSRs) in Sitka spruce (*Picea sitchensis* (Bong.) Carr.). Mol Breed 23:349–355

Auckland LD, Bui T, Zhou Y, Williams CG (2002) Conifer microsatellite handbook. Corporate, Raleigh

Bagshaw A, Pitt J, Gemmell N (2008) High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. BMC Genomics 9:49. doi:10.1186/1471-2164-9-49

Barbará T, Palma-Silva C, Paggi GM, Bered F, Fay MF, Lexer C (2007) Cross-species transfer of nuclear microsatellite markers: potential and limitations. Mol Ecol 16:3759–3767

Bérubé Y, Zhuang J, Rungis D, Ralph S, Bohlmann J, Ritland K (2007) Characterization of EST-SSRs in loblolly pine and spruce. Tree Genet Genomes 3:251–259

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping. Biotechniques 20:1004–1010

Chagné D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera MT, Vendramin GG, Garcia V, Frigerio J-M, Echt C, Richardson T, Plomion C (2004) Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor Appl Genet 109:1204–1214

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. J Hered 101:789–79. doi:10.1093/jhered/esq069

Deemer DL, Nelson CD (2010) Standardized SSR allele naming and binning among projects. Biotechniques 49:835–836

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38

Devey M, Sewell MM, Uren TL, Neale DB (1999) Comparative mapping in loblolly and radiata pine using RFLP and microsatellite markers. Theor Appl Genet 99:656–662

Echt CS, May-Marquardt P (1997) Survey of microsatellite DNA in pine. Genome 40:9–17

Echt CS, Vendramin GG, Nelson CD, Marquardt P (1999) Microsatellite DNA as shared genetic markers among conifer species. Can J For Res 29:365–371

Echt CS, Saha S, Krutovsky KV, Wimalanathan K, Erpelding JE, Liang C, Nelson CD (2011) An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. BMC Genet 12:17

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445

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

Elsik CG, Williams CG (2001) Low-copy microsatellite recovery from a conifer genome. Theor Appl Genet 103:1189–1195

Fisher PJ, Gardner RC, Richardson TE (1996) Single locus microsatellites isolated using 5′ anchored PCR. Nucleic Acids Res 24:4369–4371

González-Martínez SC, Robledo-Arnuncio JJ, Collada C, Díaz A, Williams CG, Alía R, Cervera MT (2004) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theor Appl Genet 109:103–111

Guevara MA, Chagné D, Almeida MH, Byrne M, Collada C, Favre JM, Harvengt L, Jeandroz S, Orazio C, Plomion C, Ramboer A, Rocheta M, Sebastiani F, Soto A, Vendramin GG, Cervera MT (2005) Isolation and characterization of nuclear microsatellite loci in *Pinus pinaster* Ait. Mol Ecol Notes 5:57–59

Kirkman TW (1996) Statistics to use. url: http://www.physics.csbsju.edu/stats/. Accessed 20 Dec 2010.

Kofler R, Schlötterer C, Lelley T (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23:1683–1685

Kutil BL, Williams CG (2001) Triplet-repeat microsatellites shared among hard and soft pines. J Hered 92:327–332

Lamoureux D, Peterson DG, Li W, Fellers JP, Gill BS (2005) The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). Genome 48:1120–1126

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet 109:361–369

Martin J-F, Pech N, Meglécz E, Ferreira S, Costedoat C, Dubut V, Malausa T, Gilles A (2010) Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing. BMC Genomics 11:560

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. doi:10.1371/journal.pone.0011212

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

NCBI (2010) National Center for Biotechnology Information, Taxonomy Browser for *Pinus tadea*. http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=Pinus%20taeda. Accessed 20 Dec 2010

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

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. doi:10.1186/1471-2164-11-180

Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12:795–807

Pfeiffer A, Olivieri AM, Morgante M (1997) Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome 40:411–419

Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martiensenn RA (2005) Differential methylation of genes and repeats in land plants. Genome Res 15:1431–1440

Resnick R, Stein LD (1995) STS_Pipeline1.2. http://www.broadinstitute.org/ftp/distribution/software/STS_Pipeline1.2/. Accessed 20 Dec 2010

Rousset F (2007) GENEPOP'007: a complete reimplementation of the GENEPOP software for Windows and Linux. Mol Ecol Res 8:103–106

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

Shepherd M, Cross M, Maguire TL, Dieters JM, Williams CG, Henry RJ (2002) Transpecific microsatellites for hard pines. Theor Appl Genet 104:819–827

Soranzo N, Provan J, Powell W (1998) Characterization of microsatellite loci in *Pinus sylvestris* L. Mol Ecol 7:1260–1261

Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, Sommanas W, Tragoonrung S, Srinives P (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). BMC Plant Biol 9:137. doi:10.1186/1471-2229-9-137

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

Van de Ven WTG, McNicol RJ (1996) Microsatellites as DNA markers in Sitka spruce. Theor Appl Genet 93:613–617

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotech 23:48–55

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. Science 324:1213–1216

von Stackelberg M, Rensing S, Reski R (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. BMC Plant Biol 6:9. doi:10.1186/1471-2229-6-9

Yin TM, Zhang XY, Gunter LE, Li SX, Wullschleger SD, Huang MR, Tuskan GA (2009) Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome. New Phytol 181:498–503

Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y, Sun X, Tang K (2006) Conservation of noncoding microsatellites in plants: implication for gene regulation. BMC Genomics 7:323. doi:10.1186/1471-2164-7-323

Zhou Y, Bui T, Auckland LD, Williams CG (2002) Undermethylated DNA as a source of microsatellites from a conifer genome. Genome 45:91–99