SHORT COMMUNICATION

Characterization of masson pine (*Pinus massoniana* Lamb.) microsatellite DNA by 454 genome shotgun sequencing

Tian-Dao Bai · Li-An Xu · Meng Xu · Zhang-Rong Wang

Received: 6 June 2013 / Revised: 15 November 2013 / Accepted: 19 November 2013 / Published online: 4 December 2013

Abstract Microsatellites (simple sequence repeats, SSRs) are important genetic markers in tree breeding and conservation. Here we utilized high-throughput 454 sequencing technology to mine microsatellites from masson pine (MP) genomic DNA. First, we analyzed the characteristics of SSRs in all nonredundant MP reads (genome survey sequences, GSSs) and compared them with loblolly pine (LP) GSSs and BACs (bacterial artificial chromosome clone sequences), and three other nonconiferous species GSSs. Second, a set of MP GSS–SSR primer pairs were designed. There were extremely low overall GSS–SSR densities (28 SSR/Mb) in MP when compared with LP (48 SSR/Mb) and the other species. AT, AAT, AAAT, and AAAAAT were the richest motifs in di-, tri-, tetra-, and hexanucleotides, respectively. Two hundred forty GSS–SSR primer pairs were designed in total, and 20 novel polymorphic markers were identified using three populations (two natural and one clonal seed orchard) as evaluating samples. These markers should be useful for future MP population genetics studies.

Keywords 454 genome sequencing · GSS–SSR · SSR frequency · *Pinus massoniana*

Abbreviations

GSS	Genome survey sequence
Mb	Megabase
SSR	Simple sequence repeat
BAC	Bacterial artificial chromosome clone sequence
MP	Masson pine
LP	Loblolly pine

Introduction

Masson pine (*Pinus massoniana* Lamb., MP) is not merely a major timber and wood pulp species but also an ecologically important one in montane forest ecosystems. This species grows extensively throughout across central and southern China, spreading some 12° in latitude and 21° in longitude (21°41′-33°56′N, 102°10′-123°14′E) (Zhou 2000; Qin et al. 2003). It accounts for one half of the forest stock volume there (Xu and Wang 2001). Of high fiber quality, MP is widely used in construction, furniture, and pulp industries. It is indigenous, grows fast, even under conditions too harsh for other species, and is commonly planted as a pioneer species for reforestation (Zhou 2000). Even back in the early 1980s, a national improvement program was initiated with the goal of meeting logging and pulp industry demands (Ding et al. 2005; Zhou et al. 1997). However, in terms of population genetics, there have been no marker-assisted selection (MAS) studies for MP because of the lack of effective molecular markers. Furthermore, the collection and conservation of genetic germplasm resources are important issues for MP long-term breeding. The informative DNA markers are powerful high-resolution tools for evaluation of genetic variation at the molecular level.

Microsatellites, or simple sequence repeats (SSRs) have distinguishing features such as reproducibility, abundant polymorphism, codominant inheritance, and even distribution along chromosomes. These make them a predominant tool

Communicated by P. Ingvarsson

Electronic supplementary material The online version of this article (doi:10.1007/s11295-013-0684-y) contains supplementary material, which is available to authorized users.

T.-D. Bai · L.-A. Xu (✉) · M. Xu · Z.-R. Wang
Key Laboratory of Forest Genetics and Gene Engineering,
Nanjing Forestry University, Nanjing 210037, China
e-mail: laxu@njfu.edu.cn

for population genetics, molecular ecology, and MAS studies (Li et al. 2002; Selkoe and Toonen 2006; Oliveira et al. 2006). Conventional methods for developing SSR markers in nonmodel organisms involve the construction of small–insert genomic libraries and subsequent enrichment of clones containing tandem repeat motifs (Powell et al. 1996; Kalia et al. 2011). Those methods are laborious and difficult to use. To date, only 20 MP genome SSR markers have been published (Guan et al. 2011; Hung et al. 2012), in part, probably because pine species have extremely large genomes, and thus less effort has been put into microsatellite identification. An alternative method of SSR mining is to transfer existing markers from related species, though some publications have reported available transferability with low SSR polymorphism in conifer species (Bérubé et al. 2003; Echt et al. 1999; Gonzalez-Martinez et al. 2004; Fu and Shi 2005; Mariette et al. 2001). Ai et al. (2006) tested 116 SSRs from other *Pinus* species (*P. taeda* and *P. strobus*), and only 6.90% of them were polymorphic in MP. EST-based SSRs have relatively high cross-transferability owning to the conserved nature of protein-coding regions (Lesser et al. 2012; Chagné et al. 2004; Decroocq et al. 2003; Liewlaksaneeyanawin et al. 2004). Another advantage is the genic-SSRs probably affecting gene functions and regulation (Kalia et al. 2011). Nevertheless, a main potential issue is that EST–SSRs have generally lower polymorphism levels than those of genomic SSRs. Many studies have reported a higher polymorphism and resolution in genome-derived SSRs than in EST-derived ones (Song et al. 2012; Pandey et al. 2012; Yang et al. 2005; Chabane et al. 2005; Wen et al. 2010; Ueno et al. 2012; Chagné et al. 2004). Therefore, the development of species-specific genomic SSRs should be useful for fingerprinting and parentage analysis in closely related lines. Besides, previous reports have demonstrated a higher GSS–SSR density than EST-derived ones in loblolly pine (*P. taeda*, LP) (Echt et al. 2011). If SSR markers have similar properties in MP, mining SSRs from GSSs rather than ESTs may be more practicable, especially if working on a limited budget. Fortunately, the advent, extensive utilization, and declining cost of high-throughput sequencing technologies have allowed for fast advances in the sequencing of nonmodel organisms, in particular, for organisms with large genomes such as conifers (Ritland 2012; Magbanua et al. 2011; Fernández-Pozo et al. 2011; Kovach et al. 2010; Keeling et al. 2011; Parchman et al. 2010).

In this study, we investigated and analyzed the MP microsatellites DNA obtained from 454 GS-FLX shotgun sequencing, and compared them with the GSS–SSRs from three nonconifer species, to explore the feasibility of GSS–SSR development. We then designed a set of PCR primer pairs based on the GSSs obtained to evaluate MP–SSR marker amplification and polymorphism.

Materials and methods

DNA extraction, sequencing, and assembly

Fresh MP needles were collected from a superior tree (No. 125) in a seed orchard located in Fujian province, one of the main MP production areas in China. The Cetyl–trimethyl ammonium bromide (CTAB) method (Porebski et al. 1997) with minor modifications was used for total genomic DNA extraction. Five micrograms of genomic DNA were cut into 1000–1400 bp fragments with a Sonic Dismembrator 550 (Fisher Scientific, US) and purified with Agencourt AMPure beads (Agencourt Bioscience, US). The purified DNA was then prepared as a single strand library using a GS DNA Library Preparation Kit (Roche Applied Science, US), and bound on beads using a GS emPCR Kit (Roche Applied Science, US), and eventually sequenced on a Roche 454 GS-FLX platform. A quarter run was performed, and a total of 218,871 reads with an average length of 580 bp were identified. After adaptor removal and assemblage by Newbler version 2.3 (Roche Applied Science, USA) with default parameters (minimum overlap length=40, minimum overlap identity=90%, contig length threshold=100 bp), we got 21,122 contigs and 106,428 (48.6% of total reads) singletons.

SSR mining and analysis

Only nonredundant singletons (Table 1) were utilized for SSR mining and analysis owing to the fact that the extremely large conifer species genome (C. 20–40 Gb) (Kovach et al. 2010) may result in unreliable assembly. A total 5,393 GSS sequences and 101 LP BAC clone sequences were downloaded from http://www.ncbi.nlm.nih.gov/dbGSS and http://www.ncbi.nlm.nih.gov/nuccore, respectively to compare SSR frequency and characteristics with those of the MP genome shotgun sequences. To further understand the GSS–SSR characteristics, the GSSs of the three nonconifer species, namely *Populus trichocarpa* (9.5 Mb), *Vitis vinifera* (136.7 Mb), and *Oryza sativa* (250.6 Mb) were also downloaded from National Center for Biotechnology Information (NCBI) to investigate their SSR densities and proportions. The LP BACs were also used as a reference and compared to LP and MP GSSs because of their sequence length.

The MISA program (http://pgrc.ipk-gatersleben.de/misa/misa.html) was used for SSR searches. The definition of *unit size/minimum number of repeats* were set as *2/9*, *3/6*, *4/5*, *5/5*, and *6/4*, and 0 bp was set as the interruption of two adjacent SSRs. All recovered microsatellites were classified by type (di-, tri-, tetra-, etc.) and motif (AT/TA, AG/CT, CG/GC, etc.) (Bérubé et al. 2007). Chi-square (χ^2) tests were used to compare observed SSR frequencies on various unit types between the GSSs of species according to Echt et al. (2011).

Table 1 Number and total size (bp) of sequences examined, average seq. length (bp), average G+C content, total number of SSRs, No. of SSR-containing seq., frequency of SSRs, and No. of c-SSRs in five species

Characteristics of sequences examined	*Pinus massoniana* (GSSs)	*Pinus taeda* (GSSs)	*P. taeda* (BACs)	*Populus trichocarpa* (GSSs)	*Vitis vinifera* (GSSs)	*Oryza sativa* (GSSs)
No. of sequences examined	106,428	5,393	101	11,777	229,272	408,438
Total size of examined sequences (Mb)	60.17	2.52	11.85	9.54	136.68	250.64
Average sequence length (bp)	565	468	117,308	810	596	614
Average GC content (%)	37.52	37.87	38.02	35.88	37.09	44.24
Total number of SSRs	1,699	122	758	816	9,451	10,674
No. of SSR-containing sequences	1,530	110	99	726	8,324	9,368
SSRs/Mb	28	48	64	86	69	43
No. of c-SSRs	56	4	44	25	253	179

GSSs, Genome survey sequences; *BACs,* Bacterial artificial chromosome clone sequences; *Mb,* Megabase; *c-SSR,* SSR present as compound form

SSR marker evaluation

Two wild populations and clonal seed orchard population were employed for marker evaluation. The wild populations were Shanghang (SH) and Wuping (WP) located at 25°18′N, 116°52′E, in an elevation of 970–1280 m; and 25°09′N, 115°55′E, in an elevation of 352–472 m, the orchard was Wuyi clonal seed orchard (WCS) located at 24°58′N, 117°27′E, in an elevation 434 m. DNA was isolated from the needles of 143 samples via the method mentioned above (Porebski et al. 1997) and quantified by NanoDrop 2000 (Thermo Scientific, US).

The 10 μl PCR reaction mixture containing 50 ng DNA, 10 mM Tris–HCl, 2.5 mM MgCl2, 0.25 mM dNTP, 0.3 μM primer, and 0.5 U Taq polymerase (Takara, Dalian, China) was run on a GeneAmp® PCR system 9700 (Applied Biosystems, US). The PCR program was as follows: 5 min denaturation at 94 °C, followed by 30 cycles of 30 s at 94 °C, 30 s at 50/51 °C (see Table 1), and 30 s at 72 °C, then a final extension at 72 °C for 5 min. PCR products were run on an 8 % denaturing polyacrylamide gel and the bands in each lane were carefully recorded.

Data analysis was performed on Popgene version 1.32 (Yeh et al. 1999), including the number of alleles per locus (*Na*), observed heterozygosity (*Ho*) and expected heterozygosity (*He*), probability of deviation from Hardy–Weinberg equilibrium (*HWE*), and calculation of linkage disequilibrium (*LD*) between all loci. The polymorphic information content (*PIC*) were calculated by PowerMarker version 3.2.5 (Liu and Muse 2005), and Micro-Checker version 2.2.3 (van Oosterhout et al. 2004) was used to estimate the frequencies of null alleles for each locus.

Results

Microsatellite identification and analysis

In total, 106,428 unique MP GSSs with an average length of 565 bp were analyzed after discarding the redundant sequences. Of these, 1,699 SSRs conforming to the definitions (i.e., *unit/minimum number of repeats 2/9*, *3/6*, *4/5*, *5/5*, *6/4*) were recovered from 1,530 sequences (1.4 % of the total sequences). Compound SSRs accounted for 3.2 % and the average SSR density was 28 SSR/Mb (Table 1). Among all of the MP GSS unit types, dimers were dominant (59.74 %), in which the AT/TA motif occurred at the highest frequency (61.67 %) (Fig. 1), followed by AG/CT (19.90 %), and AC/GT (11.01 %). AAT/ATT was the most abundant trimer (25.66 %) motif (58.49 %), and AAG/CTT was second at 14.91 %. In the tetramers (7.65 %), AAAT/ATTT (52.31 %) was the most common, and AAAAT/ATTTTT (18.82 %) was the most common hexamer (5.00 %). However, of the pentamers (1.94 %), AGCCT/AGGCT (18.18 %) and AAAAC/GTTTT (18.18 %) were the most common motifs; AAAAT/ATTTT (12.12 %) and ACTCT/AGAGT (12.12 %) were the second most.

Of the 5,393 LP GSSs (covering 2.5 Mb), a total of 122 SSRs were recovered from 110 sequences (2.0 %) with a density of 48 SSRs/Mb, which also contained 3.2 % compound SSRs. Of the 101 BACs (covering 11.8 Mb), 99 possessed microsatellites, and there were 758 SSRs (containing 5.8 % compound forms) in total with an average density of 64 SSRs/Mb (Table 1). Notably, MP GSSs exhibited a significantly lower SSR frequency than those of both LP GSSs (χ^2=64.79, p<0.0001) and BACs (χ^2=828.80, p<0.0001). MP also had the lowest average GSS–SSR density when compared to *O. sativa* (χ^2=256.89, p<0.0001), *V. vinifera* (χ^2=1,461.21, p<0.0001), and *P. trichocarpa* (χ^2=918.53, p<0.0001). Likewise, in the LP genome, dimeric SSRs were the most frequent GSSs (69.67 %) and BACs (69.79 %), followed in order by tri-, tetra-, hexa-, and pentanucleotides (Fig. 2). The most common motifs of all five unit types in BACs were AT/TA (77.50 %), AAT/ATT (61.82 %), AAAT/ATTT (56.16 %), AAAAT/ATTTT (38.46 %), and AAAAAT/ATTTTT (27.27 %). Similar results were found in GSSs, except for the pentanucleotides in which only one class of

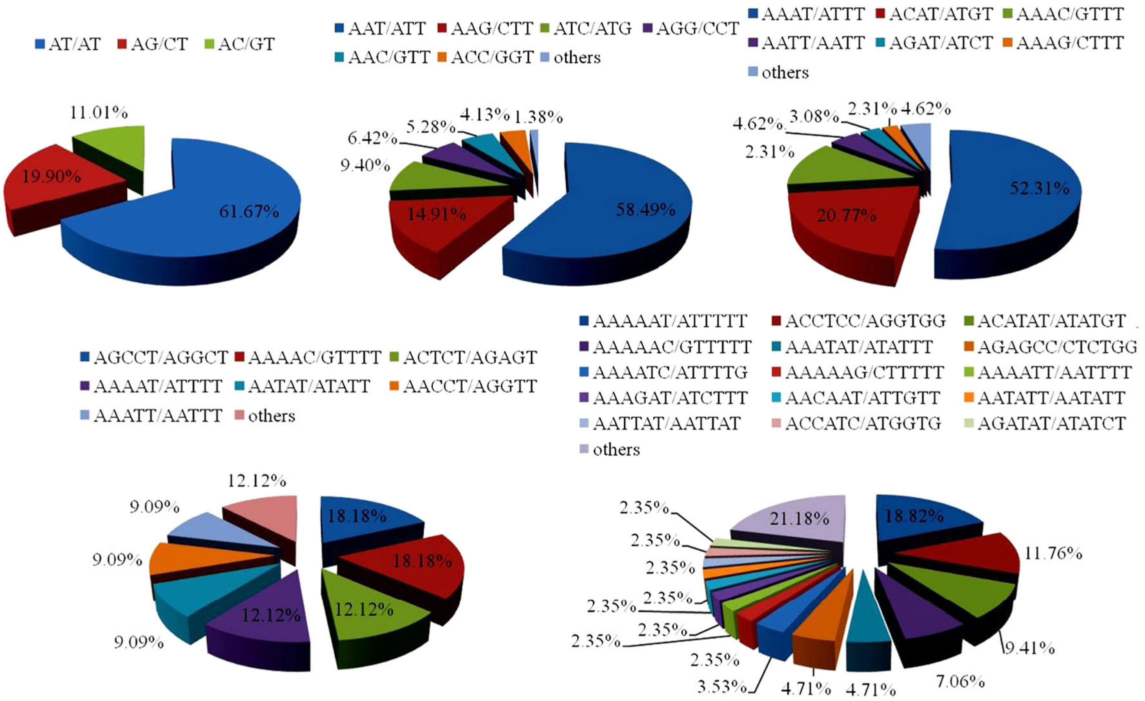

Fig. 1 Motif proportions in each *P. massoniana* unit type. Each motif in corresponding unit type showed in the form of percentage. The 'others' denotes types other than specified

motif (AAGAT/ATCTT) was identified. Interestingly, chi-square tests revealed that the dimer proportions between LP GSSs and BACs (χ^2=0.0007, p=0.98) were closer than those of MP and LP GSSs (χ^2=4.69, p=0.03). Besides, variation

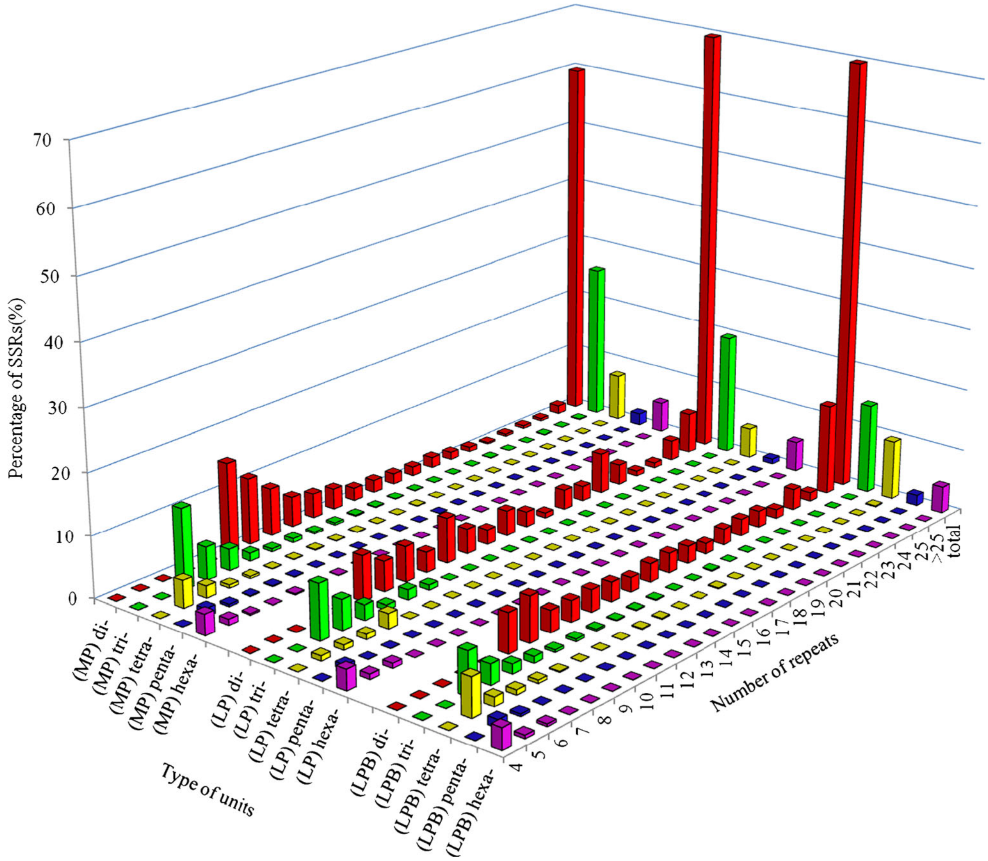

Fig. 2 Percentages of various unit types with different numbers of repeats in masson pine (MP), loblolly pine (LP) genome survey sequences, and loblolly pine BAC (LPB) clone sequences. The di-, tri-, tetra-, penta-, and hexa- denote the corresponding nucleotides unit types of microsatellites

tendency of each unit type against repeats (especially dimers) differed slightly between MP and LP (more of a fluctuating decline with the increasing number of repeats in LP BACs and GSSs than in MP GSSs) (Fig. 2), although significant Pearson's correlations were observed not only between LP BACs and GSSs ($r=0.67$, $p=0.0005$) but also between LP and MP GSSs ($r=0.63$. $p=0.0014$).

In comparing the MP GSS–SSR properties with those of the nonconifer species, similar trends in the proportions of unit types were found in *P. trichocarpa* and *V. vinifera*. Dimers were the dominant type and accounted for 59.44 % in *P. trichocarpa* and 55.75 % in *V. vinifera*, followed in order by tri-, tetra-, hexa-, and pentamers. In *O. sativa*, triplet SSRs (53.61 %) replaced dinucleotides (32.86 %) as the most common type. In terms of the motifs of the five unit types, similar to MP and LP, AT/TA, AAT/ATT, and AAAT/ATTT were the predominant motifs in *P. trichocarpa* and *V. vinifera*. But in *O. sativa*, apart from AT/TA taking first place with a percentage of 44.74 %, AAT/ATT and AAAT/ATTT were not the most abundant motifs.

Microsatellite evaluation

In total, 240 primer pairs (containing 104 di-, 91 tri-, 20 tetra-, 4 penta-, 15 hexamers, and 6 compound SSRs) were designed using Primer3 version 4.0.0 (Koressaar and Remm 2007) with a preferred amplicon size of 100–400 bp, GC content of 30–70 %, T_m value of 50–60 °C, and primer lengths of 18–25 bp. Prescreening all primer pairs with one individual randomly chosen from a wild population (SH), we successfully amplified 121 SSRs (50 %, containing 47 di-, 51 tri-, 9 tetra-, 1 penta-, 11 hexamers, and 2 compound forms) with the expected sizes, and subsequently checked in six samples. Twenty (16.5 %) SSRs (Table 2) with obvious and repeatable polymorphic bands were evaluated in the SH, WP, and WCS populations. Of these, dimers were the most abundant (9 of 20), followed by six hexamers, four trimers, and one compound trimer (Table 2). No tetrameric or pentameric loci displayed polymorphism in the candidates. An online BLASTn search showed that all of these polymorphic loci were novel and no significantly similar nucleotide sequences were found on NCBI.

The polymorphic levels of each primer pair in three populations are listed in Table 3. The mean *Na* ranged from 3.3 to 3.5, *Ho*/*He* ranged from 0.472/0.533 to 0.487/0.540, and *PIC* ranged from 0.527 to 0.534. 8 of the 20 loci deviated from *HWE* expectations in two wild populations, and three of them (P. Ma22, P. Ma96 and P. Ma116) showed evidence for null alleles in the form of observed heterozygosity deficiency (Table 3). Though a significant *LD* was detected in a few pairs of loci (e.g., P. Ma22 vs 95 ($p=0.038$) in SH; P. Ma24 vs 27 ($p=0.017$) in WP) in one or two populations, no locus pair synchronously revealed significant *LD* in all of the three populations, implying that all 20 loci are not closely linked to each other. The *Ho*/*He*/*PIC* values in the clonal seed orchard (WCS) were similar to the two wild populations, indicating that genetic diversity has been maintained in the seed orchard.

Discussion

SSRs have emerged as an important marker system in many areas of molecular genetics because they exhibit a high degree of allelic diversity. The identification and development of SSR markers represent significant challenges for nonmodel or extremely large genome size organisms. We examined the characteristics of microsatellites contained in 2–6 bp unit types of MP based on next-generation genome sequencing technology. Our work revealed an extremely low GSS–SSR frequency in MP when compared with LP and the three nonconifer species. We inferred that the MP genome has a naturally low SSR density, though the actual mechanism causing this phenomenon remains unclear. The low GSS–SSR density observed in MP compared with the other species is likely due to evolutionary divergence among the species rather than due to sequencing technique or sampling error biases. Lower GSS–SSR density in MP and LP compared with *V. vinifera* and *P. trichocarpa* partially agrees with previous reports that the occurrence of SSRs in conifers is low (Bérubé et al. 2007; Echt et al. 2011; Ueno et al. 2012), that means that using high-throughput sequencing technology in large-scale SSR mining of MP and other conifer species is a preferable or even necessary approach. It should be pointed out that, in LP, the higher SSR density in BACs than that in GSSs may imply that the frequency of GSS–SSR underrepresented the true density of genome wide SSR, because the longer and relatively complete BAC sequences may better represent the true characteristics of the LP genome.

The AT motif occurred at a higher proportion than any other dinucleotide motifs in all the analyzed species. Furthermore, the frequency of the A/T-rich motif in tri-, tetra-, penta-, and hexa- unit also took the highest or relatively higher rank in all the species except for *O. sativa* (see Table 1). This possibly demonstrates that the AT/TA motif does not only universally appear in the pine genome (Victoria et al. 2011), but also in other tree species (Ueno et al. 2012; Pootakham et al. 2012; Tuskan et al. 2004). The GSS–SSR characteristics among these species were partly compatible with the phylogenetic history of these species. The genome of *P. trichocarpa* and *V. vinifera* (xylophyta) may undergo similar environmental selection pressures to MP and LP, while *O. sativa*, a herbaceous crop, experiences different selection forces (Li et al. 2000).

Among the five categories of hexa- (73 %) and trinucleotides (56 %) seem to have higher amplification capabilities,

Table 2 Forward (F) and reverse (R) primer sequence, repeat motif, expected amplicon size, optimized annealing temperature (Ta), and GenBank accession No. of 20 *Pinus massoniana* microsatellite loci

Locus	Primer sequences(5′-3′)	Repeat motif	Size (bp)	Ta (°C)	GenBank accession no.
P. Ma22	F: GGAGAAGCCACAAGAAAGG R: CGGTAAATAAGGTTAGAGCC	$(TC)_{11}$	244	51	KC146070
P. Ma24	F: GGACCGTGGACTAGCATA R: TAGGGAATTTGGAGTTTGTG	$(TG)_{10}$	378	51	KC146071
P. Ma25	F: AAACCTCCTTCGCCTTTG R: GGATTCATTTCGCATTCATA	$(ATC)_8(TTC)_8$	165	51	KC146072
P. Ma27	F: GCATAGTCACATTGGGTTTA R: GTCAAGGCATCATAGGGA	$(TGGAGG)_4$	306	51	KC146073
P. Ma32	F: ATCTTCATGGCGTTCACC R: AATCCTTTGTTGGGAGCA	$(GAGGTG)_4$	230	51	KC146074
P. Ma43	F: GCAACCTCCATATTTCACTT R: CTTTCCAATCTTCCCTTACA	$(AAT)_6$	234	51	KC146075
P. Ma45	F: TTTCTTCTTTCCCTCCCT R: TAGATCCAATGGCCTTCA	$(CCTATC)_4$	259	51	KC146076
P. Ma51	F: ACGCACGGATAAGATTGTG R: ATCAAGTTACCCTCATTTGGA	$(GAT)_8$	227	51	KC146077
P. Ma65	F: AAGGCACTCGATCTCCTC R: TGACCTGCTTCTACACCC	$(TC)_{10}$	248	51	KC146078
P. Ma66	F: TTCCATTCCCCAAAACAC R: GCCAATGAGACCCCAATA	$(TCCACC)_5$	116	51	KC146079
P. Ma72	F: AGAACCATGCCATCCAAC R: GCTTGGAAACCTGAAAACT	$(TG)_9$	231	50	KC146080
P. Ma77	F: GACCGTACAACACTCACTTGA R: CCTCTTTCCCTTGTCCTG	$(CAA)_7$	323	51	KC146081
P. Ma80	F: AAACTGGGTTATGGTCAGAGG R: CATGGACAACCAAAGGAAG	$(TTAAAT)_4$	139	51	KC146082
P. Ma86	F: AAGAAGAGTGGGTAGGGAA R: TGCTAGACTTCCAAACCC	$(TG)_{10}$	241	50	KC146083
P. Ma95	F: CTACCGATGCGATAAGGG R: ACTCGTGACTGCGACAATAC	$(AC)_9$	303	51	KC146084
P. Ma96	F: TGACCCAATAGACTCCCTC R: AGACCTATCTAAGCACAACCC	$(AC)_9$	260	51	KC146085
P. Ma116	F: AAATGAAAATTGCAGCATG R: TGGTGAGTGATTGGGGAT	$(AT)_9$	141	51	KC146086
P. Ma117	F: CTTGTGGAATGTGACTTATGG R: CTACAACATCTAAAATCCTAATCC	$(TTG)_8$	236	51	KC146087
P. Ma216	F: GGTGAATGAACCAACCCAT R: TAGGATTCTCCCTCCAGTTC	$(TG)_{11}$	299	51	KC146088
P. Ma221	F: ATTGGGAAGTTACCTCTGAA R: TCCTGGGTCTCATCTCCTT	$(TTTTGA)_5$	286	50	KC146089

The 10 μl PCR reaction mixture containing 50 ng DNA, 10 mM Tris–HCl, 2.5 mM MgCl2, 0.25 mM dNTP, 0.3 μM primer, and 0.5 U Taq polymerase (Takara, Dalian, China) was run on a GeneAmp® PCR system 9700 (Applied Biosystems, US). The PCR program was as follows: 5 min denaturation at 94 °C, followed by 30 cycles of 30 s at 94 °C, 30 s at 50/51 °C, and 30 s at 72 °C, then a final extension at 72 °C for 5 min

while of penta- (25 %), tetra- (45 %), and dinucleotides (45 %) were comparatively low. To test this statistically, we classified all five unit types into two groups and then performed a χ^2-test because fewer tetra- and penta-nucleotides were observed. Group I was trimer plus hexamer, and group II was the sum of the other three types. The test result (χ^2=4.52, p=0.034) supported our suggestion. Considering the fact that the amplification success rate of tri- and hexamer SSRs being higher than others implies that these types of SSRs may locate in conserved or coding regions, we carried out an online BLASTn and no significant similar sequence was found. Besides, we divided these 20 novel polymorphic SSRs into two groups, group I was dimer and group II was the sum of tri- and hexamer, and then performed t tests between two groups on Na, Ho, He and PIC values from three examined populations. The result showed no significant difference between two SSR groups on Na (p=0.181), Ho (p=0.981), He (p=0.138), and PIC (p=0.139). This interesting result probably indicates that we should pay more attention to hexa- and trinucleotides SSRs in subsequent MP genomic microsatellites development, though it could not be rejected that the tri- and hexamer SSRs may reside in coding regions.

Microsatellite markers generally have a high degree of cross-transferability and are an important marker source for

Table 3 Number of alleles (*Na*), observed (*Ho*) and expected (*He*) heterozygosity, polymorphic information content (*PIC*), and estimated null allele frequency for three populations of masson pine

Locus	Shanghang (SH) (N=47; 25°18′N, 116°52′E)					Wuping (WP) (N=48; 25°09′N, 115°55′E)					Wuyi clonal seed orchard (WCS) (N=48; 24°58′N, 117°27′E)				
	Na	*Ho*	*He*	*PIC*	Est. null allele freq.	*Na*	*Ho*	*He*	*PIC*	Est. null allele freq.	*Na*	*Ho*	*He*	*PIC*	Est. null allele freq.
P. Ma22	6	0.294	0.732**	0.721	0.278	4	0.304	0.540**	0.534	0.189	3	0.487	0.568**	0.561	N/A
P. Ma24	3	0.979	0.621**	0.615	N/A	5	0.958	0.644**	0.637	N/A	3	1	0.580**	0.573	N/A
P. Ma25	5	0.766	0.771**	0.763	N/A	5	0.622	0.739	0.731	N/A	6	0.771	0.725**	0.717	N/A
P. Ma27	3	0.426	0.463	0.458	N/A	3	0.298	0.303	0.3	N/A	3	0.333	0.373**	0.369	N/A
P. Ma32	2	0.468	0.503	0.498	N/A	2	0.357	0.483	0.477	N/A	2	0.575	0.497	0.492	N/A
P. Ma43	2	0.391	0.471	0.466	N/A	3	0.458	0.506**	0.501	N/A	2	0.404	0.505	0.499	N/A
P. Ma45	2	0.311	0.463	0.458	0.146	3	0.575	0.511	0.506	N/A	3	0.478	0.521	0.515	N/A
P. Ma51	3	0.37	0.495	0.489	N/A	2	0.326	0.379	0.375	N/A	4	0.356	0.506	0.501	0.149
P. Ma65	2	0.378	0.49	0.484	N/A	2	0.477	0.506	0.5	N/A	2	0.362	0.483	0.477	N/A
P. Ma66	3	0.609	0.595	0.588	N/A	3	0.605	0.618	0.611	N/A	3	0.711	0.533	0.527	N/A
P. Ma72	4	0.362	0.435	0.43	N/A	5	0.375	0.584**	0.578	0.167	4	0.404	0.661**	0.653	0.188
P. Ma77	3	0.106	0.103	0.102	N/A	3	0.021	0.239**	0.237	0.279	3	0.083	0.082	0.081	N/A
P. Ma80	4	0.796	0.661**	0.653	N/A	4	0.723	0.667**	0.66	N/A	4	0.542	0.65	0.643	N/A
P. Ma86	2	0.234	0.3	0.296	N/A	2	0.229	0.345	0.342	N/A	2	0.229	0.389**	0.385	0.173
P. Ma95	5	0.891	0.678**	0.671	N/A	5	0.66	0.519	0.513	N/A	5	0.813	0.605	0.598	N/A
P. Ma96	4	0.435	0.679**	0.671	0.169	4	0.383	0.696**	0.688	0.214	3	0.333	0.566**	0.561	0.191
P. Ma116	3	0.085	0.518**	0.513	0.350	3	0.208	0.470**	0.465	0.246	3	0.311	0.658**	0.651	0.247
P. Ma117	3	0.83	0.641**	0.634	N/A	4	0.875	0.672**	0.664	N/A	4	0.708	0.650**	0.644	N/A
P. Ma216	4	0.644	0.698**	0.691	N/A	4	0.575	0.693**	0.686	N/A	4	0.563	0.623	0.617	N/A
P. Ma221	3	0.362	0.487**	0.482	N/A	4	0.404	0.543**	0.537	0.114	2	0.146	0.500**	0.495	0.305
Mean	3.3	0.487	0.54	0.534		3.5	0.472	0.533	0.527		3.3	0.48	0.534	0.528	

N, population sample size; ** indicates significant ($p < 0.01$) departure from Hardy–Weinberg equilibrium; N/A denotes no detectable null alleles genetically related species. Previous studies reported 54 % to 94.2 % EST–SSR transfer rates in a few *Pinus* species (Lesser et al. 2012; Chagné et al. 2004). However, the high marker transferability usually accompanies low polymorphism. Although a large percentage of EST–SSR markers were able to amplify in related species, some of them usually have very rare or even monomorphic alleles (Liewlaksaneeyanawin et al. 2004; Lesser et al. 2012; Echt et al. 2011). Instead, the species-specific genomic SSR markers generally have higher polymorphic levels than EST-based ones (Chagné et al. 2004; Liewlaksaneeyanawin et al. 2004), which is useful for fingerprinting and pedigree reconstruction of closely related individuals. According to the criteria of Botstein et al. (1980), more than half of the 20 novel polymorphic SSRs demonstrated highly informative scores ($PIC>0.50$), and the others were reasonably informative scores ($0.50>PIC>0.25$) except for the SSR of P. Ma77 ($PIC<0.25$). This means that these markers can be considered as an important marker resource in MP population genetic studies. Of these, 4–6 loci exhibited null alleles (estimated null allele frequencies range from 0.114 to 0.350) in examined populations and especially for P. Ma96 and P. Ma116, null alleles were detected in all three populations, implying that caution should be exercised when using these loci in population genetics.

Genetic improvement of MP has so far mainly involved traditional selection and crossing based on phenotypes (Ding et al. 2005; Zhou et al. 1997). The SSR markers we developed and others (Guan et al. 2011; Hung et al. 2012) can be used in fingerprinting of superior clones, paternity analysis and pedigree reconstruction of breeding populations. Thus, we expect to enhance the accuracy and progess of the MP breeding program significantly. Furthermore, our subsequent development of large-scale GSS- and EST-based SSR markers development should greatly facilitate the QTL mapping, MAS, and comparative genomics studies.

Acknowledgments This work was supported by the National High Technology Research and Development Program of China (2006AA100109) and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). We thank two anonymous reviewers for their insightful comments on previous versions.

Data Archiving Statement The twenty novel SSR sequences (GenBank accession No.: KC146070- KC146089) were uploaded on NCBI (http://www.ncbi.nlm.nih.gov/nuccore/?term=microsatellite+sequences+pinus+massoniana);

References

Ai C, Xu L-A, Lai H-L, Huang M-R, Wang Z-R (2006) Genetic Diversity and Paternity Analysis of a Seed Orchard in *Pinus massoniana*. Scientia silvae sinicae 42(11):146–150

Bérubé Y, Ritland C, Ritland K (2003) Isolation, characterization, and cross-species utility of microsatellites in yellow cedar (*Chamaecyparis nootkatensis*). Genome 46(3):353–361. doi:10.1139/g03-014

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(3):251–259. doi:10.1007/s11295-006-0061-1

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32(3):314

Chabane K, Ablett GA, Cordeiro GM, Valkoun J, Henry RJ (2005) EST versus genomic derived microsatellite markers for genotyping wild and cultivated barley. Genet Resour Crop Evol 52(7):903–909. doi:10.1007/s10722-003-6112-7

Chagné D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera M, Vendramin G, Garcia V, Frigerio J, Echt C (2004) Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor Appl Genet 109(6):1204–1214

Decroocq V, Fave M, Hagen L, Bordenave L, Decroocq S (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. Theor Appl Genet 106(5):912–922

Ding G, Zhou Z, Wang Z (2005) Cultivation and Utilization of Pulpwood Stand for *Pinus massoniana*, 1st edn. China Forestry Publishing House, Beijing, China

Echt CS, Saha S, Deemer DL, Nelson CD (2011) Microsatellite DNA in genomic survey sequences and UniGenes of loblolly pine. Tree Genet Genomes 7(4):773–780. doi:10.1007/s11295-011-0373-7

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

Fernández-Pozo N, Canales J, Guerrero-Fernández D, Villalobos DP, Diaz-Moreno SM, Bautista R, Flores-Monterroso A, Guevara MÁ, Perdiguero P, Collada C, Cervera MT, Soto Á, Ordás R, Canton FR, Avila C, Cánovas FM, Claros MG (2011) EuroPineDB: a high-coverage web database for maritime pine transcriptome. BMC Genomics 12:366. doi:10.1186/1471-2164-12-366

Fu X-X, Shi J-S (2005) Identification of seeds of pinus species by microsatellite markers. J For Res 16(4):281–284

Gonzalez-Martinez SC, Robledo-Arnuncio JJ, Collada C, Diaz A, Williams CG, Alia R, Cervera MT (2004) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theor Appl Genet 109(1):103–111. doi:10.1007/s00122-004-1596-x

Guan L, Suharyanto SS (2011) Isolation and characterization of tetranucleotide microsatellite loci in *Pinus massoniana* (Pinaceae). Am J Bot 98(8):e216–217. doi:10.3732/ajb.1100076

Hung K-H, Lin C-Y, Huang C-C, Hwang C-C, Hsu T-W, Kuo Y-L, Wang W-K, Hung C-Y, Chiang T-Y (2012) Isolation and characterization of microsatellite loci from *Pinus massoniana* (Pinaceae). Bot Stud 53:191–196

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. Euphytica 177(3):309–334. doi:10.1007/s10681-010-0286-9

Keeling CI, Weisshaar S, Ralph SG, Jancsik S, Hamberger B, Dullat HK, Bohlmann J (2011) Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (Picea spp.). BMC Plant Biol 11:43

Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10):1289–1291

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, Neale DB (2010) The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11:420–433. doi:10.1186/1471-2164-11-420

Lesser MR, Parchman TL, BUERKLE C (2012) Cross-species transferability of SSR loci developed from transciptome sequencing in lodgepole pine. Mol Ecol Resour 12(3):448–455

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

Li Y-C, Röder MS, Fahima T, Kirzhner VM, Beiles A, Korol AB, Nevo E (2000) Natural selection causing microsatellite divergence in wild emmer wheat at the ecologically variable microsite at Ammiad, Israel. Theor Appl Genet 100:985–999

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(2):361–369. doi:10.1007/s00122-004-1635-7

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9):2128–2129

Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, Cronn RC, Nelson CD, Peterson DG (2011) Adventures in the Enormous: A 1.8 Million Clone BAC Library for the 21.7 Gb Genome of Loblolly Pine. PLoS ONE 6(1):e16214

Mariette S, Chagné D, Decroocq S, Vendramin GG, Lalanne C, Madur D, Plomion C (2001) Microsatellite markers for Pinus pinaster Ait. Ann For Sci 58(2):203–206

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. Genet Mol Biol 29(2):294–307

Pandey MK, Gautami B, Jayakumar T, Sriswathi M, Upadhyaya HD, Gowda MVC, Radhakrishnan T, Bertioli DJ, Knapp SJ, Cook DR, Varshney RK (2012) Highly informative genic and genomic SSR markers to facilitate molecular breeding in cultivated groundnut (*Arachis hypogaea*). Plant Breed 131(1):139–147. doi:10.1111/j.1439- 0523.2011.01911.x

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: 180–195

Pootakham W, Chanprasert J, Jomchai N, Sangsrakru D, Yoocha T, Tragoonrung S, Tangphatsornruang S (2012) Development of genomic- derived simple sequence repeat markers in *Hevea brasiliensis* from 454 genome shotgun sequences. Plant Breed 131(4):555–562. doi:10.1111/j.1439-0523.2012.01982.x

Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Report 15(1):8–15

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. TRENDS in Plant Sci 1(7):215–222

Qin G, Zhang J, Jing G, Zou S, Tong Q (2003) Geographical provenance of *Pinus massoniana*, 1st edn. Zhejiang University Press, Zhe Jiang

Ritland K (2012) Genomics of a phylum distant from flowering plants: conifers. Tree Genet Genomes 8(3):573–582. doi:10.1007/s11295-012-0497-4

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett 9(5): 615–629. doi:10.1111/j.1461-0248.2006.00889.x

Song Y-P, Jiang X-B, Zhang M, Wang Z-L, Bo W-H, An X-M, Zhang D-Q, Zhang Z-Y (2012) Differences of EST-SSR and genomic-SSR markers in assessing genetic diversity in poplar. Forestry Studies in China 14(1):1–7. doi:10.1007/s11632-012-0106-5

Tuskan GA, Gunter LE, Yang ZK, Yin T, Sewell MM, DiFazio SP (2004) Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*. Can J For Res 34(1):85–93. doi:10.1139/x03-283

Ueno S, Moriguchi Y, Uchiyama K, Ujino-Ihara T, Futamura N, Sakurai T, Shinohara K, Tsumura Y (2012) A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, *Cryptomeria japonica*. BMC Genomics 13:136. doi:10.1186/1471-2164-13-136

van Oosterhout C, Hutchinson WF, Wills DP, Shipley P (2004) Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4(3):535–538

Victoria FC, da Maia LC, de Oliveira AC (2011) In silico comparative analysis of SSR markers in plants. BMC Plant Biol 11:15. doi:10.1186/1471-2229-11-15

Wen M, Wang H, Xia Z, Zou M, Lu C, Wang W (2010) Developmenrt of EST-SSR and genomic-SSR markers to assess genetic diversity in *Jatropha Curcas* L. BMC Res Notes 3:42. doi:10.1186/1756-0500-3-42

Xu L-A, Wang Z-R (2001) Index and strategy in different variation levels on selection breeding for pulpwood in masson pine. J Nanjing For Univer 25(2):19–22

Yang X, Liu P, Han Z, Ni Z, Sun Q (2005) Gentic diversity revealed by genomic-SSR and EST-SSR markers among common wheat, spelt and compactum. Prog Nat Sci 15(1):24–33

Yeh FC, Yang RC, Boyle T (1999) POPGENE VERSION 1.31, Microsoft Window-based freeware for population genetic analysis, quick user guide

Zhou Z (2000) Masson Pine in China, 1st edn. China Forestry Publishing House, Beijing

Zhou Z, Qin G, LI G, Huang G, Lan Y, Zhong D (1997) Achievements, problems and its countermeasures of genetic improvement of masson pine. For Res 10(4):435–442