Survey of microsatellite DNA in pine

C.S. Echt and P. May-Marquardt

Abstract: A large insert genomic library from eastern white pine (*Pinus strobus*) was probed for the microsatellite motifs $(AC)_n$ and $(AG)_n$, all 10 trinucleotide motifs, and 22 of the 33 possible tetranucleotide motifs. For comparison with a species from a different subgenus, a loblolly pine (*Pinus taeda*) genomic library was also probed with the same set of di- and tri-nucleotide repeats and 11 of the tetranucleotide repeats. The four most abundant microsatellite motifs in both species were $(AC)_n$, $(AG)_n$, $(AAT)_n$, and $(ATC)_n$, which as a group accounted for over half the microsatellite sites investigated. The two dinucleotide repeats were the most abundant microsatellite motifs tested in both species, each at 2–4.5 sites/megabase pair (Mbp), but the two trinucleotide motifs were nearly as abundant and are considered good candidates for pine microsatellite marker development efforts. Eastern white pine had more than twice as many $(AC)_n$ as $(AG)_n$ loci, in contrast with loblolly pine and most other plant species in which $(AG)_n$ is more abundant. In both pine species the minimum estimated genome density for all microsatellites, excluding $(AT)_n$ repeats, was 16 sites/Mbp.

Key words: *Pinus strobus*, *Pinus taeda*, eastern white pine, loblolly pine, simple sequence repeats.

Résumé : Une banque génomique à inserts de grande taille provenant du pin blanc (*Pinus strobus*) a été criblée avec les motifs microsatellites $(AC)_n$ et $(AG)_n$, avec chacun des 10 motifs trinucléotidiques et avec 22 des 33 motifs tétranucléotidiques possibles. Pour des fins de comparaison entre espèces appartenant à des sous-genres différents, une banque génomique du *Pinus taeda* (« loblolly pine ») a également été criblée avec le même jeu de sondes à motifs di- et tri-nucléotidique ainsi qu'avec 11 des sondes à motif tétranucléotidique. Les quatre motifs les plus abondants chez les deux espèces étaient $(AC)_n$, $(AG)_n$, $(AAT)_n$ et $(ATC)_n$, lesquels comptaient pour plus de la moitié des sites étudiés. Les motifs dinucléotidiques étaient les plus abondants chez les deux espèces puisqu'on trouvait 2 à 4,5 sites/Mpb pour chacun de ces motifs. Les deux motifs trinucléotidiques étaient presque aussi abondants et constituent de bons candidats pour le développement de marqueurs microsatellites chez le pin. Le pin blanc possédait plus de deux fois plus de loci $(AC)_n$ que de loci $(AG)_n$, contrairement au *P. taeda* et la plupart des autres espèces végétales chez lesquelles le motif $(AG)_n$ est plus abondant. Chez les deux espèces de pin, la densité génomique minimale pour tous les microsatellites, en excluant $(AT)_n$, a été estimée à 16 sites/Mpb.

Mots clés : *Pinus strobus*, *Pinus taeda*, pin blanc, microsatellites.

[Traduit par la Rédaction]

Introduction

Simple sequence repeats (SSRs), also known as microsatellites or short tandem repeats, have proven valuable as genetic markers because they are abundant, variable, and easy to genotype (Tautz and Renz 1984; Tautz 1989; Litt and Luty 1989; Weber and May 1989). Allele sizes differ by the length of the repeat unit and codominant markers can be amplified by polymerase chain reaction (PCR) from small amounts of DNA. These features make SSRs desirable markers for genome mapping, tree improvement breeding, or population genetics studies and they have led us to begin marker development in two pine species: eastern white pine (*Pinus strobus*), of the subgenus *Strobus*, and loblolly pine (*Pinus taeda*), of the subgenus *Pinus* (Echt et al. 1996).

Corresponding Editor: J.P. Gustafson.

Received July 22, 1996. Accepted October 7, 1996.

C.S. Echt[1] and P. May-Marquardt. United States Department of Agriculture, Forest Service Research, North Central Forest Experiment Station, 5985 County Road K, Rhinelander, WI 54501, U.S.A.

[1] Author to whom all correspondence should be addressed (e-mail: cecht@newnorth.net).

SSR marker development is currently a lengthy and expensive process, especially for species with large genomes such as pines, in which 2C ≅ 50 pg (Wakamiya et al. 1993). Whether markers are obtained from SSR-enriched libraries or not, it is advantageous to select the most abundant SSR motifs for marker development. There are 47 di-, tri-, and tetra-nucleotide repeat motifs (Jin et al. 1994), but only a few are present at high frequency in any given organism. For example, the most abundant SSR in several well-studied mammals is $(AC)_n$ (Beckmann and Weber 1992; Rothuizen et al. 1994), while in many plant species, $(AT)_n$ or $(AG)_n$ is the most abundant SSR (Lagercrantz et al. 1993; Morgante and Olivieri 1993; Wang et al. 1994). Notable differences in SSR frequencies are found among plant genera, however, and the factors controlling SSR frequencies and distributions are not known. A database search can reveal the most abundant SSRs in genera that have ample DNA sequence information (Morgante and Olivieri 1993; Wang et al. 1994), but for *Pinus* few suitable DNA sequences are available, making it necessary to estimate SSR frequencies from genomic libraries by hybridization methods. Comparisons of frequency estimates obtained from mammalian database and genomic library surveys show good agreement between the two approaches (Beckmann and Weber 1992; Stallings et al. 1991). Southern hybridization of SSR probes to restriction-digested DNA can define relative SSR

abundance within or between species (Poulsen et al. 1993; Smith and Devey 1994; Depeiges et al. 1995), but the method cannot generate frequency estimates.

In the current study, genomic frequencies of 34 SSRs in eastern white pine and 23 SSRs in loblolly pine were estimated by probing large insert genomic libraries with biotin-labeled oligonucleotide SSRs. Di-, tri-, and tetra-nucleotide repeats, but no self-complementary repeats, were surveyed. Our goals were to learn whether soft pine and hard pine genomes were organized similarly with respect to various repeat sequences and to find which repeat motifs were most abundant and, therefore, most suitable, for SSR marker development efforts.

Materials and methods

Plant material and library construction

DNA from megagametophytes of eastern white pine, clone P-312, was purified by cesium chloride gradient centrifugation. The DNA was partially digested with *Mbo*I, and then Klenow polymerase was used in a partial fill-in reaction to give the restriction fragments 3′-AG-5′ overhangs (Sambrook et al. 1989). Fragments of 14–23 kilobase pairs (kbp) were purified on an agarose gel and then ligated into complementary partially filled in *Xho*I sites of the phage vector λFix II (Stratagene, Inc.). Ligation products were assembled into phage particles with Gigapack II XL packaging extract and the resulting phage was used to transfect XL1Blue, MRA(P2) *Escherichia coli* cells. Ligation, packaging, and transfection were performed according to the manufacturer's instructions (Stratagene, Inc.). The resulting genomic library had an estimated insert size of 20 kbp. Use of restriction-minus host cells allowed cloning of methylated genomic pine DNA, while the presence of the P2 phage in host cells insured that only recombinant phage would develop into plaques. The XL-type of packaging extract allowed preferential cloning of large inserts. Control libraries, made without insert DNA, generated no plaques, showing a practical absence of uncut or religated vector in the library.

A genomic library of loblolly pine, clone 7-56, was kindly provided by John Davis, University of Florida—Gainesville, U.S.A. This library was constructed essentially in the same way as that for eastern white pine, except the vector used was λGEM-12 and the phage particles were assembled using Packagene extract (Promega, Inc.), resulting in an average insert size of 15 kbp. Estimations of SSR frequencies were only made from the primary unamplified libraries from each species.

Probe synthesis, calculation of T_d, and nomenclature

The following 5′-biotinylated oligonucleotide probes were synthesized on an Applied Biosystems 391 DNA synthesizer: $(AG)_{15}$, $(CA)_{15}$, $(GT)_{15}$, $(AAG)_{10}$, $(ACG)_{10}$, $(ATT)_{10}$, $(CGG)_{10}$, $(CTA)_{10}$, $(CTC)_{10}$, $(GTG)_{10}$, $(GTT)_{10}$, $(TCA)_{10}$, $(TGC)_{10}$, $(AAAC)_{10}$, $(AAAG)_{10}$, $(ACCC)_8$, $(ACTT)_8$, $(ATCC)_8$, $(CAAT)_8$, $(CCGT)_8$, $(CCTT)_{10}$, $(CGGG)_{10}$, $(CGTT)_8$, $(GAAT)_8$, $(GCGT)_8$, $(GGGA)_8$, $(GGTC)_{10}$, $(GGTT)_{10}$, $(GTTA)_8$, $(TACC)_{10}$, $(TATC)_{10}$, $(TGAG)_8$, $(TTGC)_8$, and $(TTTA)_{10}$. In addition, two alkaline-phosphatase-labeled probes, $(CA)_{14}$ and $(ACAG)_{7.75}$, were used according to the manufacturer's instructions (FMC, Inc., U.S.A.). The temperature of dissociation, T_d, for each probe was obtained using the base composition function of the MACVECTOR v.4.5.3 DNA sequence analysis program (Oxford Molecular Group, Campbell, Calif., U.S.A.). T_d is calculated to account for double-strand dissociation from membrane-bound DNA and is 7.6°C lower than the T_m (Rychlik and Rhoads 1989). For oligonucleotides that are 30 bases or fewer, MACVECTOR employs the nearest-neighbor method (Rychlik and Rhoads 1989), while for longer oligonucleotides, a base composition equation is used (Baldino et al. 1989). SSR motifs are named using the convention of alphabetical ordering (Tables 1 and A1) and the motif name is always used unless a specific probe sequence is referred to.

Plaque lifts, hybridizations, and probe detection

Phage plaque DNA was transferred and UV cross-linked to 132 mm diameter nylon membranes (Immobilon-S, Millipore Corp.) using NEBlot Photope protocols (New England Biolabs). Duplicate or triplicate membrane transfers were made from each plate. To eliminate nonspecific chemiluminescent signals resulting from endogenous biotinylated proteins present in the plaques, plaque-lift membranes were treated with proteinase K (50 μg/mL, 0.01 M Tris–Cl (pH 8.0), 0.005 M NaEDTA, plus 0.5% SDS) for 1 h at 50°C and washed at room temperature three times with 0.5× SSC (standard saline citrate: 20× = 3 M sodium chloride plus 0.3 M sodium citrate, pH 7) plus 0.5% SDS. Plaque counts for each set of phage library platings varied between 1000 and 8700 plaques per plate. Frequency estimates for abundant SSR motifs were obtained from the lower density lifts, while estimates of rare SSR motifs were obtained from the higher density lifts.

Membranes were prehybridized for 1 h in 6× SSC, 5× Denhardt's solution, plus 0.5% SDS (Sambrook et al. 1989), and were hybridized 1–16 h in 6× SSC, 0.5% SDS, plus 2 nM biotinylated oligonucleotide probe. The temperatures used for prehybridization and hybridization were -25 ± 3°C from the probe T_m. Hybridized membranes were washed twice for 5 min each with 0.5× SSC plus 0.1% SDS at room temperature, followed by two washes of 15 min each at $T_d - 3$ (±2)°C (Table 1). Hybridized probe was detected by chemiluminescence using the NEBlot Photope (New England Biolabs) procedure with minor modifications. The changes made were that four rinses or washes were used instead of two, the SDS blocking solution was filtered through Whatman 3MM paper before use, and the substrate buffer was 30 mM 2-amino-2-methyl-1-propanol (pH 9.6) plus 1 mM $MgCl_2$. Processed membranes were incubated for 2–3 h in the dark at 37°C and then exposed to x-ray film for 0.5 h at room temperature. When the filters were to be rehybridized they were stripped of probe by agitation in 0.2 mM NaOH plus 1% SDS for 20 min, rinsed three times in 0.5× SSC plus 0.1% SDS, and then air-dried.

Retrotransposon element analysis

Estimates of the number of IFG retrotransposon elements (Kossack 1989) were made using a biotinylated probe to a 4-kb fragment spanning the functional domains and one long terminal repeat of the element. The probe was labeled by nick translation (Sambrook et al. 1989) using biotin–dUTP as the label.

Data analysis

Counts of hybridizing plaques were made from digital images

Table 1. Hybridization stringency wash temperatures and ΔT_d used with various probes.

Sequence		Wash temperature[a]	ΔT_d
SSR motif	Probe	(°C)	(°C)
—	IFG	65	– 5
$(AC)_n$	$(GT)_{15}$	53	– 2
$(AG)_n$	$(AG)_{15}$	54	– 1
$(AAC)_n$	$(GTT)_{10}$	46	– 3
$(AAG)_n$	$(AAG)_{10}$	46	– 3
$(AAT)_n$	$(ATT)_{10}$	33	– 2
$(ACC)_n$	$(GTG)_{10}$	60	– 2
$(ACG)_n$	$(ACG)_{10}$	60	– 2
$(ACT)_n$	$(CTA)_{10}$	46	– 3
$(AGC)_n$	$(TGC)_{10}$	60	– 2
$(AGG)_n$	$(CTC)_{10}$	60	– 2
$(ATC)_n$	$(TCA)_{10}$	45	– 3
$(CCG)_n$	$(CGG)_{10}$	73	– 3
$(AAAC)_n$	$(AAAC)_{10}$	48	– 3
$(AAAG)_n$	$(AAAG)_{10}$	48	– 3
$(AAAT)_n$	$(TTTA)_{10}$	39	– 1
$(AACC)_n$	$(GGTT)_{10}$	57	– 4
$(AACG)_n$	$(CGTT)_8$	54	– 3
$(AACT)_n$	$(GTTA)_8$	44	– 3
$(AAGG)_n$	$(CCTT)_{10}$	58	– 3
$(AAGC)_n$	$(TTGC)_8$	53	– 4
$(AAGT)_n$	$(ACTT)_8$	45	– 2
$(AATC)_n$	$(CAAT)_8$	44	– 3
$(AATG)_n$	$(GAAT)_8$	44	– 3
$(ACAG)_n$	$(ACAG)_{7.75}$	42	– 3
$(ACCC)_n$	$(ACCC)_8$	66	– 1
$(ACCG)_n$	$(GGTC)_8$	65	– 2
$(ACCT)_n$	$(TACC)_{10}$	58	– 3
$(ACGC)_n$	$(GCGT)_8$	65	– 2
$(ACGG)_n$	$(CCGT)_8$	65	– 2
$(ACTC)_n$	$(TGAG)_8$	53	– 4
$(AGAT)_n$	$(TATC)_{10}$	48	– 3
$(AGGG)_n$	$(GGGA)_8$	69	– 2
$(ATCC)_n$	$(ATCC)_8$	54	– 3
$(CCCG)_n$	$(CGGG)_{10}$	77	– 4

[a]Washes were in 0.5× SSC.

of the x-ray film luminographs. Digitization was accomplished with a flatbed scanner connected to a Macintosh computer and image analysis was performed using National Institutes of Health (NIH, Bethesda, Md.) IMAGE software. For each SSR probe, only moderate and strong hybridization signals were counted, using a modified NIH IMAGE cell-counting macro. When only a few plaques gave hybridization signals for a particular probe, counts were made manually from the luminographs. Control platings of nonrecombinant phage typically gave background counts of 0–5 plaque-like chemiluminescent signals per membrane. Estimated numbers of SSR sites in a genome were made using the 1C (haploid) genome sizes of 26.9×10^6 kbp for white pine and 19.7×10^6 kbp for loblolly pine (Wakamiya et al. 1993).

Results

Signal detection and scoring

When counting hybridizing plaques, our goal was to include only those that gave moderate and strong signals, on the assumption that this would give a more accurate estimate of the number of longer SSRs in the genome (Iizuka et al. 1993; Stone et al. 1995). Based on theoretical calculations of T_d, hybridization conditions were chosen to detect SSR sites of more than 23 nucleotides in length, and therefore correspond to those sites most likely to be genetically informative (Weber 1990; Gastier et al. 1995). However, plaque sizes varied greatly and for many probes the hybridization signals did also, making it difficult to obtain reliable counts or comparisons when scoring the x-ray film luminographs. We found that greater counting accuracy and consistency could be obtained by analyzing digital images of the luminographs (Fig. 1). By measuring pixel densities, the image analysis software allowed us to obtain valid and reproducible counts based on hybridization signal intensities alone, regardless of plaque size.

SSR and retrotransposon frequencies

The white pine genomic library was screened with 34 of the 47 possible SSR probes, while the loblolly pine library was screened with a subset of 22 of those probes used for white pine. Both libraries were also probed with an IFG retrotransposon sequence (Kossack 1989; C. Kinlaw, personal communication) to compare SSR frequencies with those of a different type of repetitive element. We did not test for the presence of the following SSRs: $(AGTC)_n$, $(ACAT)_n$, $(ACTG)_n$, $(AGCC)_n$, $(AGCG)_n$, or $(AGGC)_n$, nor did we test for any self-complementary sequences (Table A1). For probes to the same SSR motif we found no differences in frequency estimates between biotin and alkaline phosphatase labeled probes or between probes to complementary sequences, for example, $(CA)_{15}$ versus $(GT)_{15}$.

The most and least abundant SSR motifs were the same for the two species (Fig. 2, Table 2). The four most prevalent SSRs, $(AC)_n$, $(AG)_n$, $(AAT)_n$, and $(ATC)_n$, were the same in both species, and the six least prevalent SSRs in loblolly pine were among those that were also least prevalent in white pine (Table 2). At least 74 megabase pairs (Mbp) of DNA was screened to obtain each motif frequency estimate. When counts of hybridizing plaques did not exceed those for control hybridizations, the values were reported in Table 2 as simply having fewer than 1000 repeats/genome, although no signals were detected for some of these SSRs even when 175 Mbp of DNA was screened. Triplicate hybridizations were not run for all probes, but based on the replications we did run, independent counts generally varied ±15%.

When regarded as classes, di- and tri-nucleotide repeats were present in approximately equal proportions in white pine, but in loblolly pine, trinucleotide repeats formed half and dinucleotides a quarter of all SSRs (Table 3). In both species, tetranucleotide repeats were 23–24% of all SSRs investigated and the total estimated SSR frequencies were the same. Notable differences were found, however, for individual motifs, such as $(AC)_n$, $(ACC)_n$, $(ACG)_n$, $(AGC)_n$, $(AGG)_n$, $(ACTC)_n$, and $(ATCC)_n$ (Table 2). Frequencies of the IFG retrotransposon element were similar in the two species (Fig. 2, Table 2),

Fig. 1. Examples of digital images used for counting chemiluminescent signals generated by an oligonucleotide probe hybridized to membrane-bound plaque DNA. The image on the left was digitally scanned from an x-ray film luminograph. The image on the right was generated from the scanned image using the density slice function of the NIH IMAGE program to screen out low intensity signals. A cell-counting macro was then used to estimate the number of hybridization signals. In all, 5200 plaques from a white pine genomic library were lifted onto the membrane and 177 $(ATC)_n$ hybridizing plaques were counted.

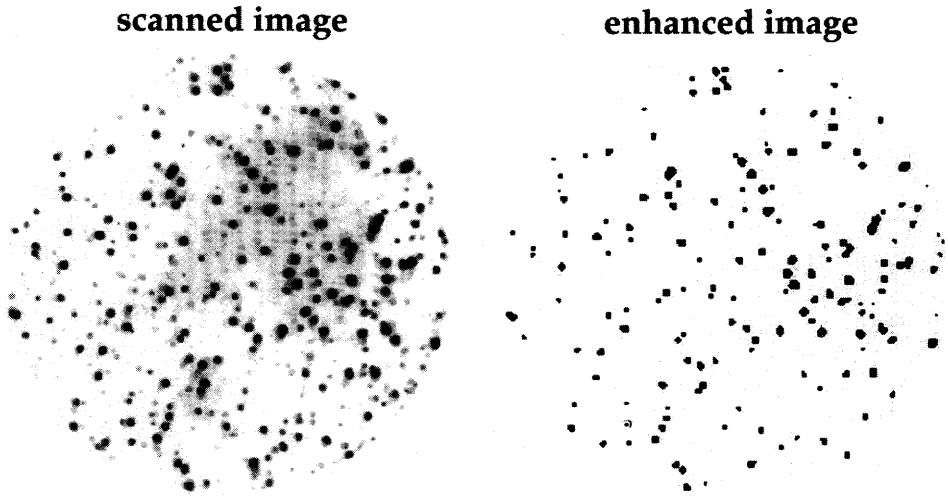
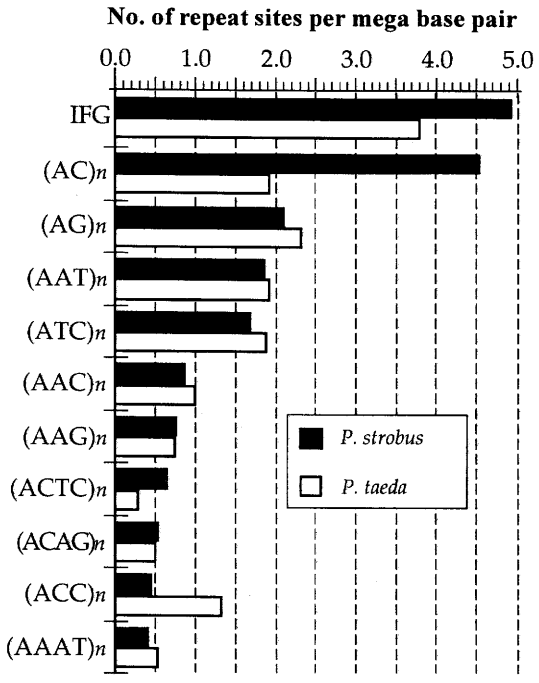


Fig. 2. Graph comparing the genome densities of abundant SSRs and the IFG retrotransposon element in eastern white pine (filled bars) and loblolly pine (open bars). The repeat motifs are ranked in order of their abundance in eastern white pine.



being equivalent to the most abundant SSR motif in white pine, $(AC)_n$, and almost twice as frequent as the most abundant SSR motif in loblolly pine, $(AG)_n$.

Discussion

The genome density for all SSRs tested was approximately the same for both species at 16–17 SSR sites/Mbp. This is a minimum estimate because some tetranucleotide SSRs were not examined. If the 11 tetranucleotide motifs not tested in white pine are each assumed to exist at the average tetranucleotide SSR frequency (4800 sites/genome), then there would be one SSR every 53 kbp, or about 19 sites/Mbp in white pine. This does not include $(AT)_n$ sites, which comprise 50% of all dicot and 37% of all monocot SSRs, as found from a database search (Wang et al. 1994). A similar extrapolation to estimate the total SSR genome density in loblolly pine was not possible, because only 11 of the 33 possible tetranucleotides (Table A1) were surveyed, but it seems reasonable to assume that the value would be similar to that of white pine. The only generalization we can make for pine concerning the relationship between SSR sequence composition and SSR frequencies is that G+C rich tetranucleotide repeats tend to be present in very low numbers, if at all (Table 2). The same cannot be said for G+C rich trinucleotide repeats, although as a class they were less abundant than the A+T rich trinucleotide SSRs.

A notable difference between the two pines was the greater abundance of $(AC)_n$ sites in eastern white pine. Surveys by hybridization (Table 4) and database searches (Akkaya et al. 1992; Morgante and Olivieri 1993; Wang et al. 1994) show that most plants have $(AG)_n$ in greater abundance than $(AC)_n$. The only other plants having more $(AC)_n$ than $(AG)_n$, among those with which similar comparisons can be made, are rice (Wu and Tanksley 1993; Panaud et al. 1995) and maize (Condit and Hubbell 1991). The factors responsible for differences in SSR abundance among plant species are not known. In primates there is evidence that certain SSRs originate from *Alu* element retrotransposition (Beckmann and Weber 1991; Arcot et al. 1995). Retrotransposons are abundant in pines

Table 2. Estimated numbers and frequencies of various repeat sequences in two pine genomes.

Sequence	10^3 sites/genome		No. of kilobase pairs between sites	
	P. strobus	*P. taeda*	*P. strobus*	*P. taeda*
IFG element	130	75	200	260
$(AC)_n$	120	38	220	520
$(AG)_n$	60	46	470	430
$(AAC)_n$	24	20	1100	970
$(AAG)_n$	21	14	1300	1400
$(AAT)_n$	51	38	530	520
$(ACC)_n$	13	27	2100	740
$(ACG)_n$	3	7	9500	2700
$(ACT)_n$	<1	<1	>58 000	>40 000[a]
$(AGC)_n$	7	10	4100	2000
$(AGG)_n$	2	5	12 000	3700
$(ATC)_n$	46	37	590	530
$(CCG)_n$	<1	<1	>58 000	>40 000
$(AAAC)_n$	8	9	3300	2200
$(AAAG)_n$	5	—	5600	—
$(AAAT)_n$	11	11	2400	1900
$(AACC)_n$	6	—	4300	—
$(AACG)_n$	9	4	2900	4600
$(AACT)_n$	7	—	3800	—
$(AAGG)_n$	<1	—	>35 000	—
$(AAGC)_n$	3	—	9500	—
$(AAGT)_n$	<1	<1	>35 000	>34 000
$(AATC)_n$	9	—	3000	—
$(AATG)_n$	6	—	4300	—
$(ACAG)_n$	14	10	1900	1900
$(ACCC)_n$	<1	—	>35 000	—
$(ACCG)_n$	<1	<1	>35 000	>34 000
$(ACCT)_n$	<1	<1	>58 000	>40 000
$(ACGC)_n$	<1	<1	>35 000	>25 000
$(ACGG)_n$	<1	<1	>35 000	>34 000
$(ACTC)_n$	18	6	1500	3200
$(AGAT)_n$	2	—	13 000	—
$(AGGG)_n$	<1	—	>35 000	—
$(ATCC)_n$	<1	3	>58 000	7700
$(CCCG)_n$	<1	—	>35 000	>40 000

Note: —, not determined.

[a]Minimum distance estimates vary depending on the total number of plaques screened.

(Kossack 1989; Kamm et al. 1996), but none have yet been identified that have an *Alu*-like structure necessary for SSR genesis as postulated by Arcot et al. (1995).

Sequences homologous to the IFG element, which contains no SSR sites (Kossack 1989), were very abundant in both eastern white pine and loblolly pine. Based on the 4-kb length of the IFG probe used, IFG sequences comprise up to 1.9 and 1.5% of each genome, respectively. These results are consistent with those for abundance of the element in *Pinus radiata*, from which IFG was originally isolated (C. Kinlaw, personal communication).

A compilation of reported frequency estimates obtained by screening genomic libraries for the two most common plant SSRs, $(AC)_n$ and $(AG)_n$, is given in Table 4. Direct comparisons of pine dinucleotide repeat frequencies with angiosperm tree species like *Piper reticulatum*, *Poulsenia armata*, *Trophis racemosa* (Condit and Hubbell 1991), or *Quercus macrocarpa* (Dow et al. 1995) were not possible because neither library insert sizes nor genome sizes were evaluated in those studies. Also not included in Table 4 is *P. radiata*, the only other pine for which frequency estimates have been reported. In that study the library was screened with a mixture of poly(AC/TG) and poly(AG/TC) probes (Smith and Devey 1994). Smith and Devey (1994) reported one dinucleotide SSR every 750 kbp, which differs from our value of one site every 230 kbp for loblolly pine (Table 3). Both species are in

Table 3. Estimated numbers and frequencies of SSR sites per haploid genome, by repeat class.

	10^3 sites/genome		No. of kilobase pairs between sites	
Class	*P. strobus*	*P. taeda*	*P. strobus*	*P. taeda*
Dinucleotides[a]	180	84	150	230
Trinucleotides	170	160	160	120
Tetranucleotides[b]	110	75	250	260
Total	450	320	60	60

[a]$(AC)_n$ and $(AG)_n$ only.

[b]Includes 22 tetranucleotide SSRs tested in *P. strobus* and 11 in *P. taeda*. The estimates for *P. taeda* have been normalized based on the assumption that the 11 SSRs not tested in *P. taeda* are present in the same proportion as they are in *P. strobus*, that is, 39% of the class total.

Table 4. Estimated dinucleotide SSR frequencies in various plant species based on hybridization surveys.

	No. of kilobase pairs between sites		
Species	$(AC)_n$	$(AG)_n$	Reference
Eastern white pine	220	470	This study
Loblolly pine	520	430	This study
Norway spruce	400	200	Morgante et al. 1996[a]
Larch	690	220	Volkaert 1995
Rice	220	480	Wu and Tanksley 1993
Wheat	290	210	Ma et al. 1996
Rapeseed	440	100	Kresovich et al. 1995
Arabidopsis thaliana	430	240	Bell and Ecker 1994

[a]M. Morgante, A. Pfeiffer, A. Costacurta, G.G. Vendramin, and A.M. Olivieri. 1996. Analysis of polymorphic simple sequence repeats in nuclear and chloroplast genomes of conifers. Poster abstract W10. The Fourth International Conference on the Plant Genome, held at San Diego, Calif., 14–18 January 1996.

section *Pinus* and would be expected to be more similar than is suggested. It may be that this difference truly reflects general variation for dinucleotide SSRs among the hard pines, but the discrepancy is more likely due to differences in library representation. The *Eco*RI λZAP II radiata pine library used by Smith and Devey had a maximum insert size of 10 kbp. Because most SSR-bearing *Eco*RI fragments are larger than 10 kbp in pine (D.N. Smith, personal communication; C.S. Echt, unpublished results), cloned $(AC)_n$ and $(AG)_n$ repeats would have been under-represented, resulting in the lower frequency estimate.

The motif $(ATC)_n$ was among the most abundant SSRs in both pine species, but appears rarely in many plants, as judged by a database survey (Wang et al. 1994). The $(AAT)_n$ repeat, on the other hand, is abundant in pines and in other plants as well (Wang et al. 1994). Compared with pines, humans have much higher frequencies of $(AC)_n$ and $(AG)_n$ repeats, with one site every 30 and 120 kbp, respectively (Beckmann and Weber 1992). Yet frequencies of $(AAC)_n$, $(AAG)_n$, $(AAT)_n$, $(ACG)_n$, $(ACT)_n$, $(AGC)_n$, and $(AGG)_n$ repeats are quite similar between humans and white pine or loblolly pine (Gastier et al. 1995). Similar genomic frequencies suggest broad phylogenetic conservation of the mechanisms that maintain certain SSR motifs, but large differences in frequencies of some SSR motifs among the two pines (see Table 2) confound speculation about general mechanisms.

In any hybridization survey the stringency of the wash conditions is critical. We found that the frequency estimates for some motifs could differ dramatically under different stringencies, as has been observed by others (Panaud et al. 1995). With the probe $(TGAG)_8$ for example, a wash at $T_d - 2°C$ resulted in no hybridization, while at $T_d - 12°C$, 80 000 sites/haploid genome were estimated. In contrast, frequency estimates with the probe $(CA)_{15}$ were only reduced by half at the higher stringency. Such differences between probes probably

reflect differences in the distribution of repeat lengths among certain SSR motifs. The hybridization wash temperature of $T_d - 3$ (±2)°C used in this study was selected to detect SSRs of at least 24 nucleotides in length, based on calculated dissociation temperatures (see Materials and methods). Thus, the values in Table 2 provide a comparison of the relative frequencies of various motifs under standardized conditions and should not be interpreted as the actual number of polymorphic SSR sites available as marker loci in pine genomes. That number is expected to be less, given no repeat variability or unsuitable flanking template sequences at some sites and given the loss of potential marker loci through cloning inefficiencies or biases. It may be, however, that some SSR motifs, CCG and CCCG, for example, may be more abundant than indicated, as a result of possible errors in the estimation of T_d (Panaud et al. 1995). There is also the issue of compound repeats, which would further reduce the expected number of genetically informative SSR loci by having two or more SSRs reside at a single locus. Compound and imperfect dinucleotide repeats are common in pines (Devey and Smith 1993; Echt et al. 1996; Fisher et al. 1996), but determination of their genomic frequencies was not possible in the present survey.

Intensive development efforts for dinucleotide microsatellites may not be warranted given the abundance of some trinucleotide repeats in pines. The advantages of tri- and tetranucleotide SSR markers are that they are easier to score in manual and automated systems because of a lower incidence of stutter bands and larger unit allelic size differences (Hauge and Litt 1993; Kijas et al. 1995; Scheffield et al. 1995). Enriched libraries based on abundant $(ATC)_n$ and $(AAT)_n$ repeats are being constructed in our laboratory for eastern white pine and loblolly pine to learn whether these trinucleotide repeats are as informative as dinucleotide SSRs. We have characterized 17 $(AC)_n$ loci from eastern white pine and find many to be highly polymorphic (Echt et al. 1996). Studies of human trinucleotide repeats show that $(AAT)_n$ loci are three times as polymorphic as $(ATC)_n$ loci (Gastier et al. 1995), and in plants, $(AAT)_n$ loci are also highly variable (Kijas et al. 1995; Rongwen et al. 1995). Development of markers for $(ATC)_n$ repeats should not be overlooked, however, since a significant portion of them, but not $(AAT)_n$ repeats, may reside in coding regions (Wang et al. 1994). The greater sequence conservation associated with coding regions would afford $(ATC)_n$ markers application among a wider range of pine species than $(AAT)_n$ markers.

References

Akkaya, M.S., Bhagwat, A.A., and Cregan, P.B. 1992. Length polymorphisms of simple sequence repeat DNA in soybean. Genetics, **132**: 1131–1139.

Arcot, S.S., Wang, Z.Y., Weber, J.L., Deininger, P.L., and Batzer, M.A. 1995. *Alu* repeats: a source for the genesis of primate microsatellites. Genomics, **29**: 136–144.

Baldino, F., Chesselet, M.-F., and Lewis, M.E. 1989. High-resolution *in situ* hybridization histochemistry. Methods Enzymol. **168**: 761–777.

Beckmann, J.S., and Weber, J.L. 1992. Survey of human and rat microsatellites. Genomics, **12**: 627–631.

Bell, C.J., and Ecker, J.R. 1994. Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*. Genomics, **19**: 137–144.

Condit, R., and Hubbell, S.P. 1991. Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. Genome, **34**: 66–71.

Depeiges, A., Goubely, C., Lenoir, A., Cocherel, S., Picard, G., Raynal, M., Grellet, F., and Delseny, M. 1995. Identification of the most represented repeated motifs in *Arabidopsis thaliana* microsatellite loci. Theor. Appl. Genet. **91**: 160–168.

Dow, B.D., Ashley, M.V., and Howe, H.F. 1995. Characterization of highly variable $(GA/CT)_n$ microsatellites in the bur oak, *Quercus macrocarpa*. Theor. Appl. Genet. **91**: 137–141.

Echt, C.S., May-Marquardt, P., Hseih, M., and Zahorchak, R. 1996. Characterization of microsatellite markers in eastern white pine. Genome, **39**: 1102–1108.

Fisher, P.J., Gardner, R.C., and Richardson, T.E. 1996. Single locus microsatellites isolated using 5′ anchored PCR. Nucleic Acids Res. **24**: 4369–4372.

Gastier, J.M., Pulido, J.C., Sunden, S., Bordy, T., Buetow, K.H., Murray, J.C., Weber, J.L., Hudson, T.J., Sheffield, V.C., and Duyk, G.M. 1995. Survey of trinucleotide repeats in the human genome: assessment of their utility as genetic markers. Hum. Mol. Genet. **4**: 1829–1836.

Hague, X.Y., and Litt, M. 1993. A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. Hum. Mol. Genet. **2**: 411–415.

Iizuka, M., Makino, R., Sekiya, T., and Hayashi, K. 1993. Selective isolation of highly polymorphic $(dG–dT)_n$ microsatellites by stringent hybridization. Genet. Anal. Tech. Appl. **10**: 2–5.

Jin, L., Zhong, Y.X., and Chakraborty, R. 1994. The exact numbers of possible microsatellite motifs. Am. J. Hum. Genet. **55**: 582–583.

Kamm, A., Doudrick, R.L., Heslop-Harrison, J.S., and Schmidt, T. 1996. The genomic and physical organization of *Ty*1-copia-like sequences as a component of large genomes in *Pinus elliottii* var. *elliottii* and other gymnosperms. Proc. Natl. Acad. Sci. U.S.A. **93**: 2708–2713.

Kijas, J.M.H., Fowler, J.C.S., and Thomas, M.R. 1995. An evaluation of sequence tagged microsatellite site markers for genetic analysis within *Citrus* and related species. Genome, **38**: 349–355.

Kossack, D. 1989. The IFG copia-like element: characterization of a transposable element present at high copy number in *Pinus* and a history of the pines using IFG as a marker. Ph.D. thesis, University of California—Davis, Davis, Calif.

Kresovich, S., Szewc-McFadden, A.K., Bliek, S.M., and McFerson, J.R. 1995. Abundance and characterization of simple-sequence repeats (SSRs) isolated from a size-fractionated genomic library of *Brassica napus* L. (rapeseed). Theor. Appl. Genet. **91**: 206–211.

Lagercrantz, U., Ellegren, H., and Andersson, L. 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. **21**: 1111–1115.

Litt, M., and Luty, J.A. 1989. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44**: 397–401.

Ma, Z.Q., Röder, M., and Sorrells, M.E. 1996. Frequencies and sequence characteristics of di-, tri-, and tetra-nucleotide microsatellites in wheat. Genome, **39**: 123–130.

Morgante, M., and Olivieri, A.M. 1993. PCR-amplified microsatellites as markers in plant genetics. Plant J. **3**: 175–182.

Panaud, O., Chen, X., and McCouch, S.R. 1995. Frequency of microsatellite sequences in rice (*Oryza sativa* L.). Genome, **38**: 1170–1176.

Poulsen, G.B., Kahl, G., and Weising, K. 1993. Abundance and polymorphism of simple repetitive DNA sequences in *Brassica napus* L. Theor. Appl. Genet. **85**: 994–1000.

Rongwen, J., Akkaya, M.S., Bhagwat, A.A., and Cregan, P.B. 1995. The use of microsatellite DNA markers for soybean genotype identification. Theor. Appl. Genet. **90**: 43–48.

Rothuizen, J., Wolfswinkel, J., Lenstra, J.A., and Frants, R.R. 1994. The incidence of mini- and micro-satellite repetitive DNA in the canine genome. Theor. Appl. Genet. **89**: 403–406.

Rychlik, W., and Rhoads, R.E. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res. **17**: 8543–8551.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. Molecular cloning: a laboratory manual. 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Scheffield, V.C., Weber, J.L., Buetow, K.H., Murray, J.C., Even, D.A., Wiles, K., Gastier, J.M., Pulido, J.C., Jandava, C., Sunden, S.L., Mattes, G., Businga, T., McClain, A., Beck, J., Scherpier, T., Gilliam, J., Zhong, J., and Duyk, G.M. 1995. A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. Hum. Mol. Genet. **4**: 1837–1844.

Smith, D.N., and Devey, M.E. 1994. Occurrence and inheritance of microsatellites in *Pinus radiata*. Genome, **37**: 977–983.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E., and Moyzis, R.K. 1991. Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. Genomics, **10**: 807–815.

Stone, R.T., Pulido, J.C., Duyk, G.M., Kappes, S.M., Keele, J.W., and Beattie, C.W. 1995. A small-insert bovine genomic library highly enriched for microsatellite repeat sequences. Mamm. Genome, **6**: 714–724.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17**: 6463–6471.

Tautz, D., and Renz, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. **12**: 4127–4137.

Volkaert, H.A. 1995. Polymorphic DNA sequences in *Larix* spp. and their use for paternity testing. Ph.D. thesis, University of Maine, Orono, Maine.

Wakamiya, I., Newton, R.J., Johnston, J.S., and Price, H.J. 1993. Genome size and environmental factors in the genus *Pinus*. Am. J. Bot. **80**: 1235–1241.

Wang, Z., Weber, J.L., Zhong, G., and Tanksley, S.D. 1994. Survey of plant short tandem DNA repeats. Theor. Appl. Genet. **88**: 1–6.

Weber, J.L. 1990. Informativeness of human $(dC–dA)_n$:$(dG–dT)_n$ polymorphisms. Genomics, **7**: 524–530.

Weber, J.L., and May, P. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44**: 388–396.

Wu, K.S., and Tanksley, S.D. 1993. Abundance, polymorphism and genetic mapping of microsatellites in rice. Mol. Gen. Genet. **241**: 225–235.

Appendix

Table A1. SSR sequence permutations: sequence motifs are ordered alphabetically in the left column; complementary permutations are to the right of the vertical bar.

Dinucleotides

AC	CA	GT	TG
AG	GA	CT	TC
AT	TA	Self-complementary	
CG	GC	Self-complementary	

Trinucleotides

AAC	CAA	ACA	GTT	TTG	TGT
AAG	AGA	GAA	CTT	TCT	TTC
AAT	ATA	TAA	ATT	TAT	TTA
ACC	CCA	CAC	GGT	TGG	GTG
ACG	CGA	GAC	CGT	TCG	GTC
ACT	CTA	TAC	AGT	TAG	GTA
AGC	GCA	CAG	GCT	TGC	CTG
AGG	GGA	GAG	CCT	TCC	CTC
ATC	TCA	CAT	GAT	TGA	ATG
CCG	CGC	GCC	CGG	GCG	GGC

Tetranucleotides

AAAC	AACA	ACAA	CAAA	GTTT	TGTT	TTGT	TTTG
AAAG	AAGA	AGAA	GAAA	CTTT	TCTT	TTCT	TTTC
AAAT	AATA	ATAA	TAAA	ATTT	TATT	TTAT	TTTA
AACC	ACCA	CCAA	CAAC	GGTT	TGGT	TTGG	GTTG
AACG	ACGA	CGAA	GAAC	CGTT	TCGT	TTCG	GTTC
AACT	ACTA	CTAA	TAAC	AGTT	TAGT	TTAG	GTTA
AAGC	AGCA	GCAA	CAAG	GCTT	TGCT	TTGC	CTTG
AAGG	AGGA	GGAA	GAAG	CCTT	TCCT	TTCC	CTTC
AAGT	AGTA	GTAA	TAAG	ACTT	TACT	TTAC	CTTA
AATC	ATCA	TCAA	CAAT	GATT	TGAT	TTGA	ATTG
AATG	ATGA	TGAA	GAAT	CATT	TCAT	TTCA	ATTC
AATT	ATTA	TTAA	TAAT		Self-complementary		
ACAG	CAGA	AGAC	GACA	CTGT	TCTG	GTCT	TGTC
ACAT	CATA	ATAC	TACA	ATGT	TATG	GTAT	TGTA
ACCC	CCCA	CCAC	CACC	GGGT	TGGG	GTGG	GGTG
ACCG	CCGA	CGAC	GACC	CGGT	TCGG	GTCG	GGTC
ACCT	CCTA	CTAC	TACC	AGGT	TAGG	GTAG	GGTA
ACGC	CGCA	GCAC	CACG	GCGT	TGCG	GTGC	CGTG
ACGG	CGGA	GGAC	GACG	CCGT	TCCG	GTCC	CGTC
ACGT	CGTA	GTAC	TACG		Self-complementary		
ACTC	CTCA	TCAC	CACT	GAGT	TGAG	GTGA	AGTG
ACTG	CTGA	TGAC	GACT	CAGT	TCAG	GTCA	AGTC
AGAT	GATA	ATAG	TAGA	ATCT	TATC	CTAT	TCTA
AGCC	GCCA	CCAG	CAGC	GGCT	TGGC	CTGG	GCTG
AGCG	GCGA	CGAG	GAGC	CGCT	TCGC	CTCG	GCTC
AGCT	GCTA	CTAG	TAGC		Self-complementary		
AGGC	GGCA	GCAG	CAGG	GCCT	TGCC	CTGC	CCTG
AGGG	GGGA	GGAG	GAGG	CCCT	TCCC	CTCC	CCTC
ATCC	TCCA	CCAT	CATC	GGAT	TGGA	ATGG	GATG
ATCG	TCGA	CGAT	GATC		Self-complementary		
ATGC	TGCA	GCAT	CATG		Self-complementary		
CCCG	CCGC	CGCC	GCCC	CGGG	GCGG	GGCG	GGGC
CCGG	CGGC	GGCC	GCCG		Self-complementary		