Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*

Gerald A. Tuskan, Lee E. Gunter, Zamin K. Yang, TongMing Yin, Mitchell M. Sewell, and Stephen P. DiFazio

Abstract: Microsatellites or simple sequence repeats (SSRs) are highly polymorphic, codominant markers that have great value for the construction of genetic maps, comparative mapping, population genetic surveys, and paternity analyses. Here, we report the development and testing of a set of SSR markers derived from shotgun sequencing from *Populus trichocarpa* Torr. & A. Gray, a nonenriched genomic DNA library, and bacterial artificial chromosomes. Approximately 23% of the 1536 genomic clones and 48% of the 768 bacterial artificial chromosome subclones contained an SSR. Of the sequences containing an SSR, 72.4% contained a dinucleotide, 19.5% a trinucleotide, and 8.1% a tetranucleotide repeat unit; 26.6% of the sequences contained multiple SSR motifs in a complex or compound repeat structures. A survey of the genome sequence database revealed very similar proportional distribution, indicating that our limited rapid, shallow sequencing effort is representative of genome-wide patterns. In total, 492 primer pairs were designed and these yielded 77 markers that were mapped in an F_2 pedigree, including 26 that were sufficiently informative to be included in a *Populus* framework map. SSRs with GC-rich motifs mapped at a significantly higher frequency than expected, although AT-rich SSRs accounted for the majority of mapped markers due to their higher representation in the genome. SSR markers developed from *P. trichocarpa* showed high utility throughout the genus, with amplification rates in excess of 70% for all *Populus* species tested. Finally, at least 30% of the markers amplified in several willow species, suggesting that some of these SSRs will be transferable across genera.

Résumé : Les microsatellites ou séquences répétées en tandem (SSRs) constituent des marqueurs codominants hautement polymorphes qui sont très utiles pour la construction de cartes génétiques, la cartographie comparée, les études de génétique des populations et les analyses de paternité. Les auteurs ont développé et testé un ensemble de marqueurs SSRs découlant du séquençage aléatoire d'une banque d'ADN génomique non enrichie et de chromosomes artificiels bactériens de *Populus trichocarpa* Torr. & A. Gray. Environ 23 % des 1536 clones génomiques et 48 % des 768 sous-clones de BAC contenaient un SSR. De ces séquences contenant un SSR, 72,4 % présentaient une répétition dinucléotidique, 19,5 % une répétition trinucléotidique et 8,1 % une répétition tétranucléotidique; 26,6 % des séquences présentaient des patrons SSRs multiples au sein de structures répétées complexes ou composées. Une étude de la base de données de la séquence du génome a révélé des proportions similaires des différents motifs, indiquant qu'un effort limité, rapide et superficiel de séquençage produit des SSRs représentatifs des patrons du génome dans son ensemble. Un total de 492 paires d'amorces ont été élaborées, résultant en 77 marqueurs qui ont été cartographiés à partir d'un pedigree F_2, incluant 26 marqueurs assez informatifs pour être inclus dans une carte de base de *Populus*. Les marqueurs SSRs à motifs riches en GC étaient cartographiés à une fréquence significativement plus élevée que la fréquence espérée, quoique les SSRs à motifs riches en AT comptaient pour la majorité des marqueurs cartographiés en raison de leur plus forte représentation dans le génome. Les marqueurs SSRs développés à partir de *P. trichocarpa* ont démontré une bonne transférabilité à travers le genre, avec des taux d'amplification excédant 70 % pour toutes les espèces de *Populus* testées. Enfin, au moins 30 % des marqueurs ont été amplifiés chez plusieurs espèces de saule, indiquant que certains de ces marqueurs SSRs seront transférables d'une genre à l'autre.

[Traduit par la Rédaction]

Introduction

Microsatellite or simple sequence repeat (SSR) markers represent a rich set of abundant, highly polymorphic codominant markers that can be used in genetic fingerprinting and clonal fidelity applications (Smulders et al. 1997; Dayanandan et al. 1998; Rajora and Rahman 2003; Wyman et al. 2003), genetic mapping and marker-aided selection procedures (Hearne et al. 1992; Echt et al. 1999; Cervera et al. 2001), and assessments of genetic diversity and phylogeny (Chase et al. 1996; Schlotterer 2001). Many types of SSRs have been shown to be preferentially associated with

Received 29 August 2003. Accepted 9 December 2003. Published on the NRC Research Press Web site at http://cjfr.nrc.ca on 14 January 2004.

G.A. Tuskan,[1] L.E. Gunter, Z.K. Yang, T. Yin, M.M. Sewell, and S.P. DiFazio. Environmental Sciences Division, P.O. Box 2008, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6422, U.S.A.

[1] Corresponding author (e-mail: tuskanga@ornl.gov).

doi: 10.1139/X03-283

transcribed regions of several plant genomes (Morgante et al. 2002) and as such may provide a means of defining functional regions of anonymous genomes. Their discovery typically involves hybridization to create SSR-enriched genomic libraries followed by sequencing of selected clones and primer design based on 5′ and 3′ flanking sequence from the microsatellite-containing fragments (Karagyozov et al. 1993). Taking advantage of the growing EST and genomic databases now publicly available, a more general approach using in silico identification, i.e., computational molecular biology, of repeats and their flanking regions has recently been explored (Toth et al. 2000; Temnykh et al. 2001; Morgante et al. 2002).

In forest tree species, SSRs have been developed for *Pseudotsuga menziesii* (Mirb.) Franco (Slavov et al. 2004), *Pinus strobus* L. (Echt et al. 1996), *Pinus radiata* D. Don. (Devey et al. 1996), *Picea sitchensis* (Bong.) Carrière (van den Ven and McNichol 1996), and eucalyptus, to name a few. In *Populus*, several hundred SSRs have been identified using various approaches for *Populus nigra* L. (van der Schoot et al. 2000; Smulders et al. 2001), *Populus tremuloides* Michx. (Dayanandan et al. 1998), and *Populus trichocarpa* Torr. & A. Gray (Frewen et al. 2000). These SSRs have been primarily derived from enriched genomic libraries and are predominantly simple, perfect di- and trinucleotide repeats of the $(GC)_n$–$(CG)_n$ and $(AG)_n$–$(TC)_n$ motifs. A motif is a reoccurring prominent sequence of repetitive DNA, including all frame shifts and complementary inverse sequences among the nucleotides within the repeat (as presented in Jurke and Pethiyagoda (1995) and Cardle et al. (2000)). Complex and compound SSRs of any motif, and AT-rich motifs in any form, are currently unavailable in public *Populus* SSR databases.

By 2004, *Populus* will be the third plant species (after *Arabidopsis* and rice) to have its genome sequenced and will represent the first tree genome to be sequenced (Wullschleger et al. 2002). As noted above, SSRs will be useful in uniting the genetic maps in *Populus* with the physical map and the sequence database. Linking each genetic map with the physical map will require the use of "framework" markers uniformly distributed across the genome. Correspondingly, SSRs associated with expressed regions will then be informative in quantitative trait locus and marker-aided selection applications.

The distribution and frequency of SSR motifs vary across intergenic, exonic, and intronic regions of genomes, with AT-rich repeats (those repeats containing two or more A and (or) T nucleotides per motif) found more often in noncoding regions and trinucleotide- and (or) GC-rich repeats (those repeats containing two or more G and (or) C nucleotides per motif) occurring more often in exonic regions (Toth et al. 2000; Temnykh et al. 2001; Morgante et al. 2002). Thus, creating a database of primer pairs for an array of all SSR motifs would potentially provide superior coverage of a genome when compared with databases containing one or a few repeat motifs. In an effort to create such a resource, we attempted to identify and develop SSRs isolated from rapid, shallow sequencing of total genomic DNA from *P. trichocarpa* and from selected bacterial artificial chromosome (BAC) clones known to contain expressed sequences (Stirling et al. 2003). The resulting SSRs were evaluated for their utility in genetic mapping in a hybrid poplar family and in interspecific amplification among members of the Salicaceae family, including *Populus* and *Salix*.

Materials and methods

Sequencing templates

Random genomic DNA

Ten micrograms of *P. trichocarpa* genomic DNA, extracted from leaf tissue of the female clone '93–968', was digested with *Pst*I (New England Biolabs, Beverly, Mass.) based on procedures by Roder et al. (1998), cloned into the *Pst*I site of pBluescript KSII+ plasmid (Stratagene, La Jolla, Calif.), and transformed into *Escherichia coli* DH5α using a standard protocol (Life Technologies, Bethesda, Md.). Plasmids were purified from bacterial cultures using the Qiaprep 96 Turbo Miniprep Kit (Qiagen, Valencia, Calif.) in preparation for sequencing. SSRs isolated from these plasmids are hereafter referred to as the "ORNL random fragment" data set.

Bacterial artificial chromosome DNA

The BACs used for this study were selected from a 10× library from *P. trichocarpa* clone 'Nisqually-1' (Stirling et al. 2001). DNA from *Populus* BACs 2c5, 12c14, 16j18, 6k8, 41g18, 46e14, 47m20, and 71j23 was purified for shotgun sequencing using the Qiagen Plasmid Mega Kit followed by two cesium chloride – ethidium bromide gradient centrifugation steps. The purified BAC DNA was sheared to an average size of 1 kb with a HydroShear (Gene Machines, San Carlos, Calif.), treated with T4 DNA polymerase (New England Biolabs) to make blunt ends, purified with QIAquick spin columns (Qiagen), and subcloned into the *Eco*RV site of pBluescript II KS+ (Stratagene). Subcloned plasmids were purified as described above. SSRs isolated from these plasmids are hereafter referred to as the "ORNL BAC fragment" data set.

Microsatellite discovery and development

Sequencing protocol

BigDye^{TM} Terminator (Applied Biosystems, Foster City, Calif.) cycle sequencing reactions were performed on a GeneAmp PCR System 9700 using modified T7 (5′-AATACGACTCACTATAGGGC-3′) or T3 (5′-AATTAACCCTCACTAAAGGG-3′) universal primers (Invitrogen, Carlsbad, Calif.), ethanol precipitated, denatured in HiDi formamide, and sequenced on an ABI PRISM 3700 DNA analyzer following the manufacturer's standard protocols (Applied Biosystems).

Sequence analysis and primer design

Base calling of the ABI trace files and assignment of quality scores were performed with Phred (Ewing et al. 1998). Sequencher (CodonCodes) was used to trim and edit the sequences. SSR motifs were identified using Finrep, a C program for finding mono-, di-, tri-, and tetranucleotide repeats (written by S. Leonardi and available at http://www.esd.ornl.gov/PGG/scripts.htm, August 2003). Primer3 (Rozen and Skaletsky 2000, available at http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi, August 2003) was used to select primer pairs flanking the SSR of interest for genotype analysis.

We screened the sequence data sets 500 bp or longer for motifs of two to four bases repeated at least four times; the minimum number of nucleotides per repeat was thus eight bases. We then placed potential repeat motifs into 14 classes: four classes for the dinucleotide repeats (i.e., $(AT)_n$–$(TA)_n$, $(AG)_n$–$(TC)_n$, $(AC)_n$–$(GT)_n$, and $(GC)_n$–$(CG)_n$) and 10 classes for the trinucleotide repeats (i.e., $(AAT)_n$–$(TAT)_n$, $(AAG)_n$–$(TCT)_n$, $(AAC)_n$–$(TGT)_n$, $(ATG)_n$–$(TCA)_n$, $(AGT)_n$–$(TAC)_n$, $(AGG)_n$–$(TCC)_n$, $(AGC)_n$–$(TGC)_n$, $(ACG)_n$–$(TCG)_n$, $(ACC)_n$–$(TGG)_n$, and $(GGC)_n$–$(CGC)_n$) based on Jurke and Pethiyagoda (1995). Because the tetranucleotide repeats were relatively rare, these repeats were considered as individual motifs and were not placed into classes. All potential repeats were also classified as (1) simple perfect, consisting of a single repeat of *n* units, (2) compound perfect, consisting of two or more alternate tandem repeats of *n* units each, or (3) complex imperfect, consisting of repeats that either (*i*) varied in motifs by a single unit (e.g., ATATATTATAT), (*ii*) consisted of alternate repeat motifs interspersed within a single region (e.g., ATATGCCGCCATAT), or (*iii*) consisted of two simple perfect motifs separated by nonrepeating sequences of variable length (e.g., ATATAT N_n GCGCGC). Primers were numbered sequentially and prefixed ORPM (Oak Ridge *Populus* microsatellite).

Microsatellite primer screening

PCR protocols and SSR products

Primers were initially screened for segregating polymorphisms using the P_1 and F_1 individuals from the interspecific hybrid F_2 family 331 derived from female *P. trichocarpa* clone '93-968' × male *P. deltoides* clone 'Ill-129' (Bradshaw and Stettler 1995). Primers passing this screen were used to genotype a subset of 44 progeny from this family. Genotyping was performed using either FAM or HEX 5′-labeled oligonucleotide primers (Operon Technologies, Inc., Alameda, Calif.). Reaction mixtures contained 25 ng of DNA, 50 ng of each SSR primer, 200 µmol/L dNTPs, 0.5 U *Taq* DNA polymerase/µL (Promega Corp., Madison, Wis.), 10 µmol/L Tris–HCl (pH 8.3), 50 mmol/L KCl, 2.0 mmol/L $MgCl_2$, 0.01% gelatin, and 0.1 mg bovine serum albumin/mL. Amplification conditions on a GeneAMP 9700 thermocycler (Applied Biosystems) included an initial denaturation step at 94 °C for 45 s followed by 30 cycles of 94 °C for 15 s, 50–55 °C for 15 s, and 72 °C for 1 min and concluded with a 5-min extension at 72 °C. Reaction products were diluted up to 1:200, denatured in HiDi formamide containing a 400-bp ROX standard (Applied Biosystems), and processed on the ABI Prism 3700 DNA analyzer. GeneScan version 3.5 was used for size calling of raw alleles based on the internal standard and Genotyper version 3.5 was used to visualize and assign alleles to categories for scoring purposes. The resulting data tables were further processed by PERL scripts (available at http://www.esd.ornl.gov/PGG/scripts.htm, August 2003) to infer inheritance of each allele. Scored data were analyzed for differences in frequency and distribution among motifs and repeat classes using a contingency χ^2 test at $\alpha \le 0.05$ and were exported for input into MapMaker (Lander et al. 1987) as a means of determining inheritance and novel mapping value.

Microsatellite mapping

Genetic segregation data from the above SSR primers and other genetic markers (provided by H.D. Bradshaw) were used to construct a genetic map from the subset of F_2 progeny from family 331. Individual male and female maps were first constructed for each F_1 using MapMaker (Lander et al. 1987). Markers were assigned to linkage groups at a logarithm of the odds minimum threshold of 5 and a maximum recombination fraction of 0.30. Using common markers, the F_1 maps were visually aligned and checked for colinearity and then integrated into a single sex-average map using JoinMap (Stam 1993). Details of map construction followed those outlined for an outbred pedigree in Sewell et al. (1999).

Species screening

The genus *Populus* consists of six sections (Eckenwalder 1996): *Abaso* Eckenwalder, *Leucoides* Spach, *Aigeiros* Duby, *Turanga* Bunge, *Tacamahaca* Spach, and *Populus* Eckenwalder. To test the utility of the SSR primers developed from *P. trichocarpa*, we selected reference species based on Eckenwalder's (1996) phylogenetic treatment. The commercial and ecological importance of reference species was also taken into consideration in our selection for this study (Table 1). Five species of *Salix* were also included to determine the general utility of the *Populus* markers in willows.

For each species, total genomic DNA from a single genotype was isolated from young leaf tips using the Qiagen DNeasy Plant Mini Kit. We randomly selected 49 mapped and 48 unmapped SSR loci for testing. Polymerase chain reaction (PCR) products were scored as dominant markers based on expected SSR size within the resolution of a 1.5% agarose gel, i.e., amplification occurred or it did not. These bands were not isolated or sequenced, and thus, SSR synteny was simply inferred. A positive control (*P. trichocarpa* clone '93-968', which yielded the expected product) and a negative control (all reaction components minus the DNA template, which yielded no product) were used for each set of reactions. Percent amplification across taxa and correlations among subgenera across all taxa were carried out on the derived data. All statistical tests were performed at $\alpha \le 0.05$.

Populus genome sequence searches

We used the FinRep program to identify microsatellites in the *Populus* genomic sequence data released by the Joint Genome Institute (JGI) (http://genome.jgi-psf.org/poplar0/poplar0.home.html, August 2003). The data were derived by shotgun sequencing of two random insert libraries (average insert size of 3 and 8 kb). The data set consisted of 4 695 113 sequence reads and 3.74 Gb of draft-quality sequence. If the *Populus* genome is between 480 and 520 Mb, then this sequencing depth represents roughly 5.0–5.5× coverage. SSRs isolated from this resource are hereafter referred to as the "JGI random fragment" data set.

Results and discussion

Microsatellite characterization

Approximately 23% of the ORNL random fragments contained at least one SSR motif that was greater than or equal to four di-, tri-, or tetranucleotide repeat units in length. A total of 1536 *Pst*I clones representing approximately 1.5 Mb (overlap unknown) of the *P. trichocarpa* genome contained

Table 1. Species of *Populus* and *Salix* used to test *Populus trichocarpa* SSR primer pair utility.

Genus, section, species	Origin	Collector, institution
Populus L.		
Aigeiros Duby		
Populus deltoides Marshall	Puyallup, Wash.	UW, WSU
Populus nigra L.	Ames, Iowa	ISU
Populus fremontii S.	Ames, Iowa	ISU
Leucoides Spach		
Populus heterophylla L.	Ames, Iowa	ISU
Populus lasiocarpa Olivier	Hube, China	NJFU
Populus Eckenwalder		
Populus tremuloides Michx.	Ames, Iowa	ISU
Populus alba L.	Xinjiang, China	NJFU
Populus grandidentata Michx.	Oak Ridge, Tenn.	ORNL
Populus tomentosa Carrière	Beijing, China	BJFU
Populus adenopoda Maxim.	Nanjing China	NJFU
Populus davidiana (Dode) Schneider	Heilongjiang, China	BJFU
Populus canescens (Ait.) Smith	Xinjiang, China	BJFU
Tacamahaca Spach		
Populus trichocarpa Torr. & A. Gray	Puyallup, Wash.	UW, WSU
Populus maximowiczii A.	Ames, Iowa	ISU
Populus yunnanensis Dode	Heilongjiang, China	BJFU
Populus simonii Carrière	Liaoning, China	BJFU
Populus catheyana Rehder	Liaoning, China	BJFU
Turanga Bunge		
Populus euphratica Olivier	Xinjiang, China	BJFU
Salix L.		
Salix suchowensis	Xuzhou, China	NJFU
Salix purpurea L.	Syracuse, N.Y.	SUNY
Salix eriocephala Michx.	Mecosta Co., Mich.	UI
Salix nigra Marsh.	Gettysburg, Pa.	UI
Salix integra Thunb.	Liaoning, China	NJFU

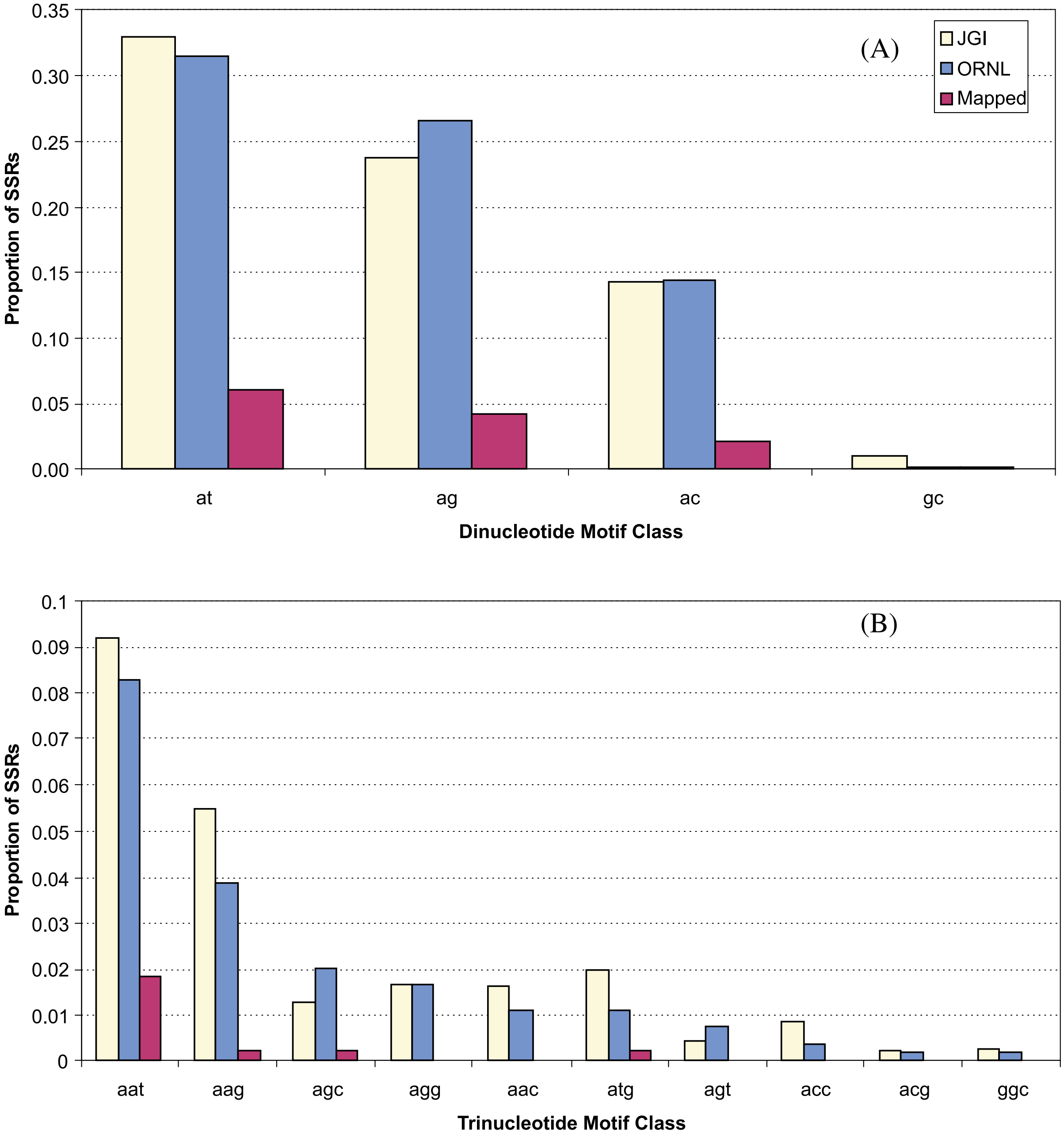
Note: NJFU, Nanjing Forestry University of China, Nanjing, China; BJFU, Beijing Forestry University of China, Beijing, China; ISU, Iowa State University, Ames, Iowa; SUNY, State University of New York, Syracuse, N.Y.; UI, University of Idaho, Moscow, Ihaho; UW, University of Washington, Seattle, Wash.; WSU, Washington State University, Pullman, Wash.

627 distinct repeats. In addition, 768 sequences from the ORNL BAC fragments, representing ~434 kb in 456 contigs with minimal overlap, contained 370 distinct repeats. Of these 997 potential SSRs, 492 had sufficient flanking sequence to design amplification primers. Of these 492, 72.4% corresponded to dinucleotide repeats, with the (AT)$_n$–(TA)$_n$ motif being the most common (31.4% of the total), (AG)$_n$–(TC)$_n$ being the next most common motif (26.5%), (AC)$_n$–(GT)$_n$ the next (14.3%), and (GC)$_n$–(CG)$_n$ being the least common motif (<1.0%) (Fig. 1A). Within the 492 SSR collection, 19.5% of the primer pairs corresponded to trinucleotide repeats and 8.1% corresponded to tetranucleotide repeats. Of the trinucleotide repeats, the (AAT)$_n$–(TAT)$_n$ motif was the most common (8.3% of the total) followed by (AAG)$_n$–(TCT)$_n$ at 3.9%, (AGG)$_n$–(TCC)$_n$ and (ACG)$_n$–(TCG)$_n$ at ~2.0%, and all others at <1.0% (Fig. 1B). The ratio among the di- and trinucleotide motifs detected in the ORNL random fragment and ORNL BAC fragment data sets is nearly identical to the ratios contained in the genome-wide JGI random fragment data set, with the exceptions of the (AG)$_n$–(TC)$_n$ repeat in the ORNL random fragment data set being slightly more abundant (26.5% versus 23.8%) and all trinucleotides being slightly less abundant (19.9% versus 23.1%) in the ORNL random fragment data set when compared with the JGI random fragment data set (Table 2; Fig. 1). Just over 70% of the repeats in the ORNL random fragment and ORNL BAC fragment data sets were simple perfect repeats; the remaining SSRs were compound perfect (6.5%) or complex imperfect (20.1%) repeats. (Note that the total number of di-, tri-, and tetranucleotide repeats is larger than the total number of SSRs because of the multiple repeat motifs found in the compound and complex classes. There were 544 distinct SSR motifs among the 492 primer pairs.)

The longest SSR in the random survey, *ORPM_224*, contained 52 repeat units in a compound perfect microsatellite, [TC]$_{18}$[AT]$_{34}$ (Table 2; Fig. 2). The mean number of repeat units across all motifs and repeat classes was seven, with the mode being four. The mean number of repeat units decreased as the size of the repeat motif increased in all databases (Table 2). Remarkably, the longest perfect repeat in the JGI data set was [AT]$_{351}$ (Table 2). A complete listing of the 492 SSR primer sequences and repeat structure in the ORNL data sets is available from the International *Populus* Genome Consortium Web site (http://www.ornl.gov/ipgc/Links.htm, August 2003).

Comparisons among studies designed to identify SSR markers are difficult to make because of differences in (*i*) marker frequency among species, (*ii*) SSR designation

Fig. 1. Variation in microsatellite occurrence among all sampled SSRs from a small-scale sequence survey (Oak Ridge National Laboratory (ORNL)), a comprehensive genomic sequencing effort (Joint Genome Institute, (JGI)), and those SSRs that were placed on a genetic map for a *Populus trichocarpa* × *Populus deltoides* F_2 pedigree. Note that the motif class includes all equivalent base repeats, e.g., AG is analogous to $(AG)_n$–$(TC)_n$ and includes AG, GA, CT, and TC repeats. (A) Dinucleotide motifs; (B) trinucleotide motifs.



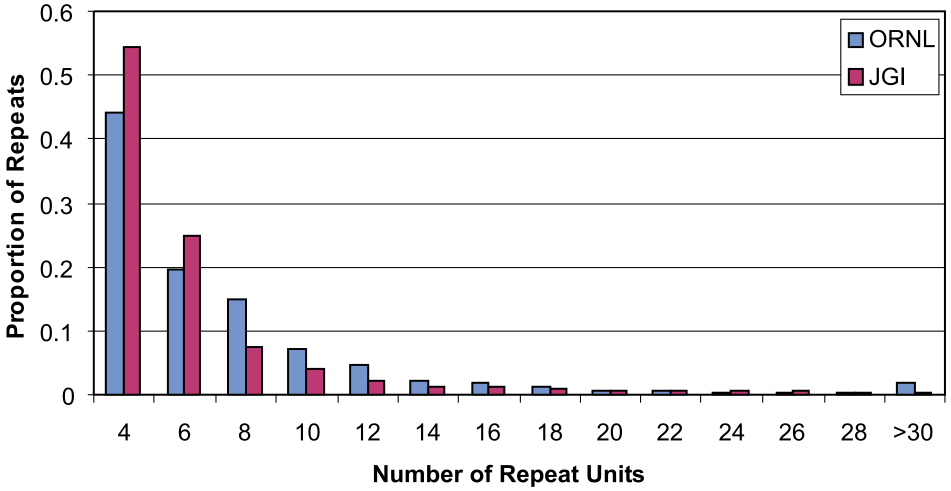
(e.g., minimum repeat numbers, inclusion of mononucleotides, etc.), and (*iii*) isolation and characterization protocols (e.g., in silico genomic and EST sequence databases versus enrichment approaches). Still, insights may be drawn by making such comparisons. Using enrichment approaches in *P. nigra*, van der Schoot et al. (2000) found that 16% of the clones in the dinucleotide library contained SSRs, whereas only 2%–7% of the clones in the mixed di- and trinucleotide libraries contained SSRs. Dayanandan et al. (1998), working with *P. tremuloides*, found that only 1% of the probed clones contained either di-, tri-, or tetranucleotide repeats. In all previous studies, the absolute frequencies of SSR occurrence are lower than those reported here using an in silico approach, although the proportional decrease from di- to tri- and tetranucleotides is similar among all *Populus* reports. The differences among studies in SSR abundance are partially because of disparities in approach and to some degree a less stringent threshold for minimum number of nucleotides in a repeat used in the current study, i.e., 24 in van der Schoot et al. (2000) and 12 in Dayanandan et al. (1998) versus 8 in the current study. Cardle et al. (2000) reported an SSR every 14 kb in the pub-

Table 2. Number and sizes of SSRs discovered in the Joint Genome Institute (JGI) and Oak Ridge National Laboratory (ORNL) sequence data sets.

	JGI, perfect repeats			ORNL, perfect repeats			ORNL, all repeats		
Motif	*N*	Mean	Max	*N*	Mean	Max	*N*	Mean	Max
Dinucleotide	223 683	6.5	351	246	7.1	38	394	7.9	52
Trinucleotide	71 745	5.1	31	102	4.9	30	106	5.9	30
Tetranucleotide	15 063	4.4	17	30	4.4	6	44	5.1	13

Note: Only perfect repeats were identified for the JGI data, whereas simple, complex, and compound repeats were characterized in the smaller ORNL data set. *N* is the total number of observed repeats (some complex repeats may be counted more than once for the comprehensive ORNL database because of complex and compound motifs of varying size), Mean is the mean number of repeat units observed, and Max is the maximum number of repeat units observed.

Fig. 2. Distribution of microsatellites across repeat units for all di-, tri-, and tetranucleotide motifs detected in *Populus trichocarpa* sequencing surveys. The Oak Ridge National Laboratory (ORNL) data are based on 493 SSRs and include both perfect and imperfect repeats, while the Joint Genome Institute (JGI) data comprise 310 491 SSRs and include only perfect repeats.



licly available *Populus* EST database available at the time of publishing. Across all motifs and repeat units, we found approximately one SSR every 4 kb of draft sequence.

In rice, Temnykh et al. (2001) reported one simple perfect SSR per 40 kb of random genomic DNA based on in silico surveys of BAC-end sequences available at the time. In examining the *Arabidopsis* EMBL database available in 2000, Cardle et al. (2000) reported an average of one SSR per 6.0 kb across all di-, tri-, and tetranucleotide repeats, with dinucleotide repeats being slightly more common than trinucleotide repeats, which were substantially more common than tetranucleotide repeats. A similar distribution pattern was reported by Katti et al. (2001). In *Arabidopsis*, the most common dinucleotide motif was $(AT)_n$–$(TA)_n$ and the most common trinucleotide motif was $(AAG)_n$–$(TCT)_n$ (Cardle et al. 2000), although this distribution varied between genomic and EST-based data sets. Morgante et al. (2002) reported a higher proportion of $(AG)_n$–$(TC)_n$ motifs in transcribed regions of the *Arabidopsis* genome, substantiating earlier claims by Toth et al. (2000), Cardle et al. (2000), and Katti et al. (2001). Across all reports, it appears that $(AT)_n$–$(TA)_n$ motifs are most frequently associated with intergenic regions of the genome and the trinucleotide repeats are found more frequently in expressed regions, with $(AAG)_n$–$(TCT)_n$ and $(AAC)_n$–$(TGT)_n$ more common in 5′ and 3′ untranslated regions. These observations are similar to the results presented here in that the SSR distributions of di- and trinucleotides for *Arabidopsis* and *Populus* are the same.

Inheritance and mapping of SSRs

All 492 primer pairs that produced PCR products were evaluated for inheritance using the parents and grandparents of an F_2 pedigree, family 331. Of these potential SSRs, 194 revealed heterozygous loci in one or both of the P_1 genotypes. Of these 194 primer pairs, 67 produced PCR products that segregated according to Mendelian expectations in the tested F_2 progeny. Note that some primer pairs produced multiple independent loci. In total, then, 77 SSR loci were placed on the *Populus* family 331 genetic map that formerly included other codominant and dominant markers. Fifty-six of these SSRs were fully informative, 6 maternally informative, 16 paternally informative, and 4 intercross informative. These 77 markers mapped across the genome on each of the 19 *Populus* chromosomes. Among these 77 markers, 26 markers contributed to the existing *Populus* framework map (Table 3). Sequences of flanking regions surrounding the repeat motif of each SSR marker placed on the framework map are available from GenBank (http://www.ncbi.nlm.nih.gov/) and are assigned accession numbers AY497381–AY497406. The proportional loss of useful SSR markers from initial sequence identification through amplification to mapped polymorphic SSR loci in this study (84%) is nearly

Table 3. *Populus trichocarpa* SSR loci, priming sequences, and repeat structures used in the framework map.

Locus	Forward primer sequence (5′–3′)	Reverse primer sequence (5′–3′)	SSR motif	Size (bp)
ORPM-015	CGTGAGTTTTGAGGCCATTT	CATGGAAAGGATCACCCACT	$[AT]_{14}$	257
ORPM-016	GCAGAAACCACTGCTAGATGC	GCTTTGAGGAGGTGTGAGGA	$[CTT]_{15}$	238
ORPM-023	ATTCCATTTGGCAATCAAGG	CCCTGAAAGTCACGTCTTCG	$[AT]_6...[AG]_6$*	197
ORPM-026	GCTGCAGTCAAATTCCAAAA	CGAGCGTCTTCTTCATGGAT	$[CA]_8$	213
ORPM-028	GGATCGACTTCCAACCCATA	AATTCCCAGATGAAGGCTCA	$[AT]_7$*	204
ORPM-029	TGGTGATCCAGTTTTGGTGA	GTCCTTGCAAGCCATGAA	$[AC]_{11}$	245
ORPM-030	ATGTCCACACCCAGATGACA	CCGGCTTCATTAAGAGTTGG	$[TC]_9$	224
ORPM-049	AAAGGGCTTTGGACGATTTT	GATTTATGAGCCTGCCCAAC	$[GA]_6$	195
ORPM-050	AAGAATTTGGGGCGGTTTAC	GCCTCAAAGGGAATTCTCAA	$[A]_7[TA]_4[A]_6$	198
ORPM-059	TGCTAGTAACTGCGCATTGG	GATGTTTTTCGCACGCATTA	$[AT]_6$	213
ORPM-064	AAAGGCCTCTGCTTCGCTAT	TTGCAGACATGATCCCAATG	$[CA]_4$	222
ORPM-127	TCAATGAGGGGTGCCATAAT	CTTTCCACTTTTGGCCCTTT	$[TG]_8$	200
ORPM-149	GTCTCTGCCACATGATCCAA	CCCGAAATGGATCAAACAAG	$[AT]_4...[CT]_4$	216
ORPM-202	TCGCAAAAGATTCTCCCAGT	TTCAAATCCCGGTAATGCTC	$[TAA]_5$	190
ORPM-203	CCACCAGGCATGAGATATGA	TCAAACCGAAAGGTCAACAA	$[TA]_4$*	209
ORPM-204	TCTGCATTGATGATTCCAGTG	GCAAGTTTTCTGCAATGTTGA	$[TC]_4$	174
ORPM-206	CCGTGGCCATTGACTCTTTA	GAACCCATTTGGTGCAAGAT	$[GCT]_7$	196
ORPM-260	TTCTAGTCCTGGCATAGCTTCA	CAGAGATTTGAATCGCAGCA	$[AAT]_{10}$	220
ORPM-276	GCAGGAGAAAACACCAGGAA	TCGCGAAAGAGAAGAAAAGC	$[TA]_6$	205
ORPM-277	CTTTGGATTGCTTGCGTTTT	TTACCATTGCTGCCATTTCA	$[GA]_4$	201
ORPM-279	TCAAATCAAACCACAAAAACACA	TGAGACGAACATATCCTTCACC	$[AT]_{18}$	197
ORPM-312	GTGGGGATCAATCCAAAAGA	CCCATATCAAACCATTTGAAAAA	$[CCT]_6$	194
ORPM-349	GAGCATGAAGCATGAGCAGA	TTTTCAGAACCAGGGGAAAA	$[AC]_{16}$	202
ORPM-381	CGGATGGATTTCATACGTGA	TGTAATTTTAGTTGAGGTTGGATTG	$[AT]_5...[AT]_4$	274
ORPM-417	TGTACCCTGCACCATCATGT	CAATTTCCAGCCCCAGTAGA	$[AT]_5$	208
ORPM-430	CCTTGGAAAAACCCCAAAAT	CAGCTCGACTCATTGCAAAA	$[AT]_9$	202

Note: All reactions were run under identical PCR as described in the Materials and methods. An asterisk indicates that the SSR motif was more complex than defined here. For instance, the sequence surrounding the *ORPM-028* motif is actually $[AT]_3[AAT]_2[AT]_7[A]_6$ in *P. trichocarpa* clone '93-968'.

identical to the average reported by Squirrell et al. (2003) for 71 plant species (83%).

Among the 77 mapped SSR markers, there were no significant differences in the frequency of di-, tri-, and tetranucleotide markers in the total SSR pool and the proportion that were placed on the genetic map (χ^2 = 1.42, df = 2, $p > 0.05$). However, a significantly higher than expected proportion of the simple perfect SSRs were placed on the genetic map when compared with the compound perfect or complex imperfect classes of markers (χ^2 = 17.96, df = 2, $p \le 0.001$). Likewise, a significantly higher than expected proportion of the SSRs with eight or more repeat units were placed on the map compared with SSRs with seven or fewer repeat units (χ^2 = 792.56, df = 1, $p \le 0.0001$). These results suggest that among the SSRs detected in this study, the simple SSRs and SSRs of seven repeat units or greater are more polymorphic in *Populus* than are the compound or shorter SSRs.

Among the dinucleotide repeats, the $(AT)_n$–$(TA)_n$ motif produced the largest number of mapped markers (Fig. 1), although the $(AG)_n$–$(TC)_n$ motif had significantly more mapped repeats than expected (χ^2 = 278.87, df = 3, $p \le 0.0001$). Similarly, among the trinucleotide repeats, the $(AAT)_n$–$(TAT)_n$ motif produced the largest number of mapped markers (Fig. 1B), whereas the $(AGC)_n$–$(TGC)_n$ motif produced significantly more mapped repeats than expected (χ^2 = 78.94, df = 9, $p \le 0.0001$). This disproportional difference among motifs was unexpected. There should have been more mapped AT-rich repeats, given that they were more abundant and have been shown to be more polymorphic. Alternatively, AT-rich repeats tend to mutate more frequently and have inherent problems during PCR that may explain their lower representation in the mapped population. Temnykh et al. (2001) have reported similar results in rice.

For the general purposes of generating genetic maps, SSRs associated with either transcribed or nontranscribed portions of the genome are equally valuable. However, for quantitative trait locus studies or marker-aided selection, SSRs within transcribed regions of the genome will have greater utility. Alternatively, SSRs associated with intergenic space will be more constructive when linking a genetic map with (*i*) physical maps or (*ii*) random draft genomic sequences where uniform, dispersed SSR distribution would be vital. The in silico approach used in this study produced microsatellites that should be useful in both general genetic mapping and physical mapping scenarios.

Utility of SSRs in other species

Because of their demonstrated inheritance and their reported value in genetic studies, we tested the capacity of a random selection of SSR primers to yield PCR products in several alternative *Populus* and *Salix* species. All tested SSR primer pairs generally displayed high frequency of amplification within and across sections at the subgenus level within *Populus*. There was particularly high amplification frequency within the *Tacamahaca*, *Leucoides*, and *Aigeros* sections, with 80%–99% of all tested primers yielding product followed by *Populus* at 70%–80% and then by *Turanga* at ~70%. In *Salix*, the frequency of amplification was predictably lower, ranging from 40% to 50%, except in *Salix integra* Thunb., in which only 30% of the *P. trichocarpa* SSR primers yielded amplified product. A χ^2 test for independence for amplification frequency among species within the same sections illustrated that species in the section *Tacamahaca* displayed significantly higher amplification than expected ($\chi^2 = 10.68$, df = 4, $p \leq 0.03$), most likely because the SSRs were derived from a member of the section, *P. trichocarpa*. The section *Populus*, represented by *P. tremuloides*, *Populus alba* L., *Populus grandidentata* Michx., *Populus tomentosa* Carrière, *Populus adenopoda* Maxim., *Populus davidiana* (Dode) Schneider, and *Populus canescens* (Ait.) Smith, showed significantly lower amplification frequency than expected. As with rice (Temnykh et al. 2001), and as has been suggested by Schlotterer (2001), amplification frequency appears to vary in concert with genetic relatedness among taxa. Therefore, it should be possible to predict the utility of these primers based on genetic distance from *P. trichocarpa*.

Conclusions

SSRs appear to be abundant in the *Populus* genome and can therefore be efficiently isolated via random shotgun sequencing. Such an approach resulted in a mix of SSR types with an accompanying variety of marker characteristics. Furthermore, our results from the ORNL BAC fragments suggest that targeting regions known to contain expressed sequences may increase the efficiency of SSR isolation, in keeping with observations that SSRs tend to be associated with low-copy DNA and coding regions in other species (Morgante et al. 2002; Toth et al. 2000; Cardle et al. 2000; Katti et al. 2001). The relatively small-scale shotgun sequencing effort (~0.004×) yielded results that were highly concordant with those obtained with comprehensive shotgun sequencing of the entire genome (~5.2×), suggesting that the microsatellite contents of other species could be well characterized with very shallow shotgun sequencing. In addition, the frequency of amplification of the discovered markers across other *Populus* species suggests that they will prove useful across the genus as well as among species from other closely related genera. As expected, the priming sites are often unique and highly conserved sequence tags, boding well for comparative mapping. Finally, these SSR markers should serve as the primary bridge connecting the genome sequence of *P. trichocarpa* to genetic improvement programs of many other tree species.

Acknowledgments

The authors acknowledge the contributions made by H.D. Bradshaw, University of Washington, for providing the TxD family 331 DNA, J. Johnson and J. Whisler, Washington State University, for providing additional plant materials for family 331, C. Loopstra and J. Vrebalov, Texas A&M University, for the construction of the *Populus* BAC library, R. Hall, Iowa State University, S. Brunsfeld, University of Idaho, R. Kopp, State University of New York, M. Huang, Nanjing Forestry University, and W. Song, Beijing Forestry University, for supplying plant tissue for species screening, D. Rokhsar for coordinating the shotgun sequencing effort at the Joint Genome Institute, J.M. Tuskan for insightful discussions regarding marker development, and M.E. Schuster, Oak Ridge National Laboratory, for help with GenBank submissions. Support for this project was provided by D. Neale, U.S. Forest Service, through a joint DOE/USDA Agenda 2020 project, the U.S. Department of Energy, Office of Science, Biological, and Environmental Research, Terrestrial Carbon Program, and the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract No. DE-AC05-00OR22725.

References

Bradshaw, H.D., Jr., and Stettler, R.F. 1995. Molecular genetics of growth and development in *Populus*. 4. Mapping QTLs with large effects on growth, form and phenology traits in a forest tree. Genetics, **139**: 963–973.

Cardle, L., Ramsey, L., Milbourne, D., Macaulay, M., Marshall, D., and Waugh, R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics, **156**: 847–854.

Cervera, M.T., Storme, V., Ivens, B., Gusmao, J., Liu, B.H., Hostyn, V., Van Slycken, J., Van Montagu, M., and Boerjan, W. 2001. Dense genetic linkage maps of three *Populus* species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. Genetics, **158**: 787–809.

Chase, M.R., Kesseli, R.V., and Bawa, K.S. 1996. Microsatellite markers for population and conservation genetics of tropical trees. Am. J. Bot. **83**: 51–57.

Dayanandan, S., Rajora, O.P., and Bawa, K.S. 1998. Isolation and characterization of microsatellites in trembling aspen (*Populus tremuloides*). Theor. Appl. Genet. **96**: 950–956.

Devey, M.E., Bell, J.C., Smith, D.N., Neale, D.B., and Moran, G.F. 1996. A genetic linkage map for *Pinus radiata* based on RFLP, RAPD and microsatellite markers. Theor. Appl. Genet. **92**: 673–679.

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

Echt, C.S., Vendramin, G.G., Nelson, C.D., and Marquardt, P. 1999. Microsatellite DNA as shared genetic markers among conifer species. Can. J. For. Res. **29**: 365–371.

Eckenwalder, J.E. 1984. Natural intersectional hybridization between North American species of *Populus* (Salicaceae) in sections Aigeiros and Tacamahaca. II. Taxonomy. Can. J. Bot. **62**: 325–335.

Eckenwalder, J.E. 1996. Systematics and evolution of *Populus*. Chap. 1. *In* Biology of *Populus* and its implications for management and conservation. *Edited by* R.F. Stettler, H.D. Bradshaw, P.E. Heilman, and T.M. Hinckley. NRC Research Press, Ottawa, Ont. pp. 7–32.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. **8**: 175–185.

Frewen, B.E., Chen, T.H.H., Howe, G.T., Davis, J., Rohde, A., Boerjan, W., and Bradshaw, H.D. 2000. Quantitative trait loci and candidate gene mapping of bud set and bud flush in *Populus*. Genetics, **154**: 837–845.

Hearne, C., Ghosh, S., and Todd, J. 1992. Microsatellites for linkage analysis of genetic traits. Trends Genet. **8**: 288–294.

Jurke, J., and Puthiyagoda, C. 1995. Simple repeat DNA sequence from primates: compilation and analysis. J. Mol. Evol. **40**: 120–126.

Karagyozov, L., Kalcheva, I.D., and Chapman, V.M. 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Res. **21**: 3911–3912.

Katti, M.V., Ranjekar, P.K., and Gupta, V.S. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. **18**: 1161–1167.

Lander, E., Abrahamson, J., Barlow, A., Daly, M., Lincoln, S., Newburg, L., and Green, P. 1987. Mapmaker a computer package for constructing genetic-linkage maps. Cytogenet. Cell Genet. **46**: 642–642.

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

Rajora, O.P., and Rahman, M.H. 2003. Microsatellite DNA and RAPD fingerprinting, identification and genetic relationships of hybrid poplar (*Populus* × *canadensis*) cultivars. Theor. Appl. Genet. **106**: 470–477.

Roder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.H., Leroy, P., and Ganal, M.W. 1998. A microsatellite map of wheat. Genetics, **149**: 2007–2023.

Rozen, S., and Skaletsky, H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. *In* Bioinformatics methods and protocols: methods in molecular biology. *Edited by* S. Krawetz and S. Misener. Humana Press, Totowa, N.J. pp. 365–386.

Schlotterer, C. 2001. Genealogical inference of closely related species based on microsatellites. Genet. Res. **78**: 209–212.

Sewell, M.M., Sherman, B.K., and Neale, D.B. 1999. A consensus map for loblolly pine (*Pinus taeda* L.). I. Construction and integration of individual linkage maps from two outbred three-generation pedigrees. Genetics, **151**: 321–330.

Slavov, G.T., Howe, G.T., Yakovlev, I., Edwards, K.J., Krutovskii, K.V., Tuskan, G.A., Carlson, J.E., Strauss, S.H., and Adams, W.T. 2004. Highly variable SSR markers in Douglas-fir: Mendelian inheritance and map locations. Theor. Appl. Genet. In press. [Published online, 19 November 2003, at http://www.springerlink.com/media/f83ebu4j80ug7v9812r2/contributions/f/e/a/2/FEA202VHC2T5KRUV_html/fulltext.html].

Smulders, M.J.M., Bredemerjer, G., Rus-Kortekaas, W., Arens, P., and Vosman, B. 1997. Use of short microsatellites from data base sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor. Appl. Genet. **94**: 264–272.

Smulders, M.J.M., van der Schoot, J., Arens, P., and Vosman, B. 2001. Trinucleotide repeat microsatellite markers for black poplar (*Populus nigra* L.). Mol. Ecol. Notes, **1**: 188–190.

Squirrell, J., Hollingsworth, R.M., Woodhead, M., Russell, J., Lowe, A.J., Gibby, M., and Powell, W. 2003. How much effort is required to isolate nuclear microsatellites from plants? Mol. Ecol. **12**: 1339–1348.

Stam, P. 1993. Construction of integrated genetic-linkage maps by means of a new computer package — JoinMap. Plant J. **3**: 739–744.

Stirling, B., Newcombe, G., Vrebalov, J., Bosdet, I., and Bradshaw, H.D., Jr. 2001. Suppressed recombination around the MXC3 locus, a major gene for resistance to poplar leaf rust. Theor. Appl. Genet. **103**: 1129–1137.

Stirling, B., Yang, Z.K., Gunter, L.E., Tuskan, G.A., and Bradshaw, H.D., Jr. 2003. Comparative sequence analysis between orthologous regions of the *Arabidopsis* and *Populus* genomes reveals substantial synteny and microcolinearity. Can. J. For. Res. **33**: 2245–2251.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and 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.

Toth, G., Gaspari, Z., and Jurke, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. **10**: 967–981.

van der Schoot, J., Pospiskova, M., Vosman, B., and Smulders, M.J.M. 2000. Development and characterization of microsatellite markers in black poplar (*Populus nigra* L.). Theor. Appl. Genet. **101**: 317–322.

van der Ven, W.T.G., and McNichol, R.J. 1996. Microsatellite as DNA markers in Sitka spruce. Theor. Appl. Genet. **93**: 613–617.

Wullschleger, S.D., Tuskan, G.A., and DiFazio, S.P. 2002. Genomics and the tree physiologist. Tree Physiol. **22**: 1273–1276.

Wyman, J., Bruneau, A., and Tremblay, M.F. 2003. Microsatellite analysis of genetic diversity in four populations of *Populus tremuloides* in Quebec. Can. J. Bot. **81**: 360–267.