

Characterization of EST-SSRs in loblolly pine and spruce

Yanik Bérubé · Jun Zhuang · Dainis Rungis ·
Steven Ralph · Jörg Bohlmann · Kermit Ritland

Received: 3 May 2006 / Revised: 14 June 2006 / Accepted: 14 August 2006
© Springer-Verlag 2006

Abstract In the first large study of conifer expressed sequence tag-simple sequence repeats (EST-SSRs), two large conifer EST databases were characterized for EST-SSRs. One database was from “interior spruce” (white and Engelmann spruce in Southern British Columbia) and Sitka spruce, while the other was from loblolly pine. We found 475 and 629 unique EST-SSRs in loblolly pine and spruce, respectively. 3′ ESTs contained 14% more SSRs than 5′ EST reads in loblolly pine and 41% more in spruce. Conifer EST-SSRs differed conspicuously from angiosperm EST-SSRs in several aspects. EST-SSRs were considerably less frequent in conifers (one EST-SSR every ~50 kb) than in angiosperms (one EST-SSR every ~20 kb). Dinucleotide repeats were the most abundant repeat class in conifers, while in angiosperms, trinucleotides were most common. Finally, the AT motif was the dominant motif recovered in both conifer species, whereas AG was the most common dinucleotide repeat in angiosperms. Also, as these EST-SSRs in conifers could be developed into useful genetic markers, our work demonstrates the value of large-scale EST sequencing projects for *in-silico* approaches for marker development.

Keywords Expressed sequence tags · Microsatellites · Simple sequence repeats · Conifers

Communicated by O. Savolainen

Y. Bérubé · J. Zhuang · D. Rungis · J. Bohlmann · K. Ritland (✉)
Department of Forest Sciences, University of British Columbia,
Vancouver, BC V6T 1Z4, Canada
e-mail: kritland@interchange.ubc.ca

S. Ralph · J. Bohlmann
Michael Smith Laboratories, University of British Columbia,
Vancouver, BC V6T 1Z4, Canada

Introduction

Microsatellites or simple sequence repeats (SSRs) are short tandemly repeated stretches of DNA composed of 1–6 bp units. They have been shown to occur in both coding and noncoding regions of all higher organisms (Tautz and Renz 1984; Gupta et al. 1996) and are thought to result from the mutational effects of replication slippage and unequal recombination (Richards and Sutherland 1992; Schlötterer and Tautz 1992; Khajavi et al. 2001). Because of their abundance and high mutation rate, they have been extremely useful as genetic markers, especially for the purposes of investigating genetic variation and relatedness and performing linkage mapping (reviewed by Jarne and Lagoda 1996; Powell et al. 1996).

In coniferous species, SSRs have traditionally been found by hybridizing repeat-enriched molecular probes in genomic libraries (e.g., van de Ven and McNicol 1996; Pfeiffer et al. 1997; Rajora et al. 2000; Hodgetts et al. 2001; Scotti et al. 2002a,b). These methods typically yield relatively few SSRs probably because of the large and repetitive genomes of conifers (Pfeiffer et al. 1997; Bérubé et al. 2003), which give complex multilocus amplification products (Echt et al. 1996; Soranzo et al. 1998).

The recent surge of interest in functional genomics has given rise to a new source of information for the development of microsatellites in plants and other organisms: expressed sequence tags (ESTs) (Scott 2000). ESTs are segments of expressed genes, captured by high-throughput single-pass sequencing of cDNA libraries (Adams et al. 1993). They contain both protein-coding sequences as well as untranslated regions (UTRs) on either end of the gene. Extensive databases of ESTs exist for many species. While these databases have been utilized in the identification of single nucleotide polymorphisms, these databases can also be used to recover SSRs

(Cardle et al. 2000; Kantety et al. 2002; Varshney et al. 2002; Gao et al. 2003).

While EST-derived SSRs (EST-SSRs) exhibit less polymorphism than genomic sequence-based SSRs, they have inherent advantages: they can be easily found, they are present in gene-rich regions of the genome, they are relatively abundant, and they are more easily transferable to related species (Scott et al. 2000; Cordeiro et al. 2001; Rungis et al. 2004). These attributes render EST-SSRs especially attractive for studies concerned with comparative genetic mapping. SSRs in gene transcripts may also play a role in gene function as, for example, trinucleotide repeats are associated with Huntington's disease in humans (The Huntington's Disease Collaborative Research Group 1993) and an SSR in the 5' UTR is associated with amylose content in rice (Ayers et al. 1997).

Although many studies have investigated the recovery of SSRs from large EST databases, only recently have conifers received attention (Chagné et al. 2004; Liewlaksaneeyanawin et al. 2004; Rungis et al. 2004). The evolutionary pattern of these EST-SSRs has not received attention. This study reports the recovery of SSRs from large EST databases in two coniferous groups: (1) spruce—composed of interior spruce (white spruce, *Picea glauca*, Engelmann spruce, *Picea engelmannii*, and their hybrids) and Sitka spruce (*Picea sitchensis*), and (2) loblolly pine (*Pinus taeda*). The three spruces exhibit small evolutionary distances, as evidenced by their ability to readily hybridize with each other (Wright 1955) and difficulties in resolving phylogenetic relationships (see Ledig et al. 2004 for a review), so we treat these species as one in this paper. We then compile the abundance and distribution of these recovered SSRs, compare their differences between spruce and pine, and, finally, compare the patterns in conifers with that previously found in angiosperms.

Materials and methods

ESTs generated by our Treenomix research team (deposited at NCBI) and ESTs from NCBI's dbEST (<ftp://ftp.ncbi.nih.gov/repository/dbEST/>) were compiled for interior spruce (composed of white spruce, Engelmann spruce, and their hybrids) and Sitka spruce. This taxonomic group is generally referred to as "spruce" in this study. ESTs for loblolly pine were obtained from NCBI. All sequence information was gathered on February 3, 2005. In both taxonomic groups, all ESTs were separated into three groups: 5'-end, 3'-end, and unknown.

All ESTs were initially searched for the presence of microsatellites using an object-oriented Java program developed in-house by Y.B. and J.Z. (SSRFinder version 2, <http://www.genetics.forestry.ubc.ca/ritland/programs.html>). Motifs

were searched for using Java's regular expressions (`java.util.regex`) and included perfect di-, tri-, tetra-, penta-, and hexanucleotide repeats as well as compound repeats composed of di-, tri-, and tetranucleotide repeats and imperfect repeats. In the perfect microsatellites, the minimum number of repeat units for dinucleotides was nine, six repeat units for trinucleotides, five repeat units for both tetra- and pentanucleotides, and four repeats for hexanucleotide repeats. These limits were chosen based on earlier studies that investigated plant EST-SSRs (Cardle et al. 2000; Tóth et al. 2000; Temnykh et al. 2001; Kantety et al. 2002; Varshney et al. 2002; Gao et al. 2003; Rungis et al. 2004). In the compound and imperfect repeats, the minimum length of di-, tri-, and tetranucleotide repeats was five units. Imperfect repeats were defined as having no more than one disruptive element of length ≥ 1 and ≤ 20 bp. The program also classified recovered microsatellites by type (e.g., di-, tri-, tetra-, etc.) and motif (e.g., AC, AG, AT, and CG) (following Jurka and Pethiyagoda 1995) where repeat motifs that were circular permutations and/or reverse complements of each other were clustered together (e.g., ACT=CTA=TCA=TGA=GAT=AGT). Counts of members within these groups were recorded.

To remove redundant comparisons, sequences that contained SSRs were assembled into contigs using the sequence assembly program CAP3 with its default parameters (Huang and Madan 1999). The SSRFinder program was then used a second time to retrieve information about repeat type and frequency. To compare between species, SSR counts were standardized by the number of nucleotides surveyed.

Results

EST resources

In spruce, 128,031 ESTs were surveyed for a total of 77.4 million base pairs (Mbp). In loblolly pine, 185,777 ESTs were surveyed for a total of 10.6 Mbp (Table 1). In spruce, the majority of ESTs (~67%) were from the 3'-end, while in loblolly pine, most (~68%) of the ESTs were from the 5'-end of the clone (Table 1). After the assembly of these ESTs, the numbers of nonredundant base pairs for spruce were 8.2, 18.1, and 6.7 Mbp for 5', 3', and unknown ESTs, respectively, while for loblolly pine, the corresponding numbers were 17.3, 10.6, and 3.8 (Table 1).

SSR-containing ESTs

In the original search (comprising all ESTs), a total of 2,048 loblolly pine ESTs and 1,416 spruce ESTs were found to contain SSRs (Table 2). These numbers correspond to 1.1% of total ESTs containing SSRs in both

Table 1 Details of EST sequence assembly in loblolly pine and spruce

Species	Direction	Number of ESTs	Number of contigs	Number of singletons	Number (bp)
Loblolly pine	5'	127,017	8,646	15,486	17,257,918
	3'	43,080	5,595	7,960	10,559,111
	Unknown	15,680	1,744	4,889	3,759,272
Spruce	5'	22,807	3,799	7,264	8,237,957
	3'	85,090	9,177	14,533	18,120,437
	Unknown	20,134	3,571	5,889	6,674,256

Number of contigs, singletons, and resulting total number of nucleotides in both species separated by EST read direction (5', 3', and unknown)

loblolly pine and spruce. After assembly of the sequences containing SSRs, these numbers were reduced to 562 in loblolly pine and 761 in spruce, a reduction of 72.6 and 46.3%, respectively. Combining 5', 3', and reads of unknown direction, 437 unique sequences were found to contain SSRs in loblolly pine while this number in spruce was 599 (both a reduction of ~22%). This complete assembly of all ESTs within the two species resulted in a nonredundant set of 475 SSRs in loblolly pine and 629 SSRs in spruce. The compilation of all SSRs reveals that in loblolly pine ESTs, on average, an SSR can be found every 56.6 kb, while in spruce ESTs, this number is lower at one SSR every 42.9 kb.

The most abundant SSR class recovered in both groups was dinucleotide repeats. These represented 35.1% of all SSRs recovered in loblolly pine and 45.2% in spruce (Fig. 1). The second most abundant SSR class recovered was trinucleotide repeats in both groups (33.6% in loblolly pine and 33.5% in spruce). Hexanucleotide repeats were the third most abundant SSR type recovered. These represented 22.9 and 13.6% of all SSRs found in loblolly pine and spruce, respectively. Tetranucleotide and pentanucleotide repeats were found in low frequencies of 4.1 and 2.0%, respectively, in loblolly pine and 3.2 and 2.4%, respectively,

in spruce. Compound and imperfect repeats were rare in both species with frequencies <2.0%.

Dinucleotide repeats The AT dinucleotide repeat motif was the most abundant motif detected in both loblolly pine (72.2% over all directions) and spruce (72.0% over all directions), and this was observed regardless of the direction of the EST read in which it was found (Table 3). This was followed by the motif AG, which represented 26.9 and 22.5% of overall dinucleotides, found in loblolly pine and spruce, respectively. The motif AC motif was rarely found in either group (0.9 and 5.5% overall in loblolly pine and spruce, respectively), and the CG motif was not observed.

Occurrences of the AT repeat motif were more frequent in 3' reads in both groups (Fig. 2a), while SSRs composed of AG and AC repeats appeared relatively evenly distributed between 5' and 3' reads in both species. Cumulative frequencies, once weighted by the number of bases in each nonredundant set of sequences in all three sequencing directions, showed that dinucleotide repeats were on average 1.4 times more frequent in 3' sequenced ESTs than in their 5' counterparts in loblolly pine and 2.3 times in spruce.

Trinucleotide repeats All eight trinucleotide repeat motif types were detected in both groups. The most abundant motif in both taxonomic groups was ACG with overall frequencies of 22.7% in loblolly pine and 21.5% in spruce (Table 3). In loblolly pine, the second most abundant motif was AAG (19.3%), while in spruce, it was AAT (19.6%). Both species shared AGG as the third most abundant trinucleotide repeat motif (12.6% in loblolly pine and 18.1% in spruce) and AAC as the least abundant of these motifs (5.3% in loblolly pine and 3.0% in spruce). A compilation of motif occurrences reveals no detectable base composition bias (A+T vs C+G) in either group.

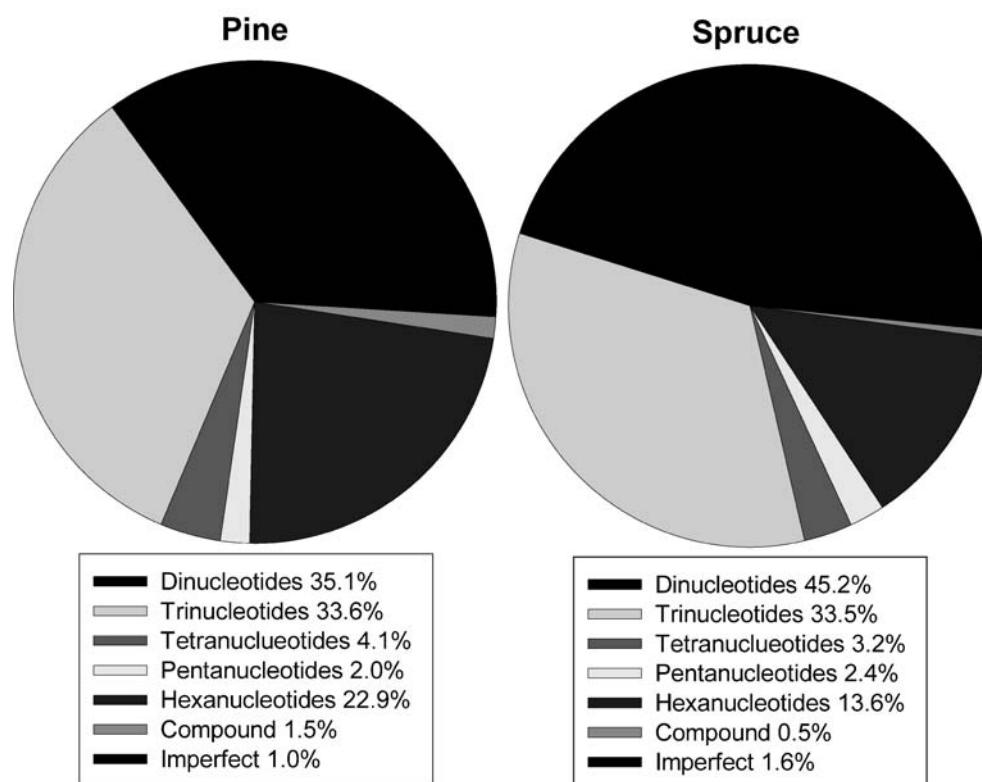
In loblolly pine, seven of the eight trinucleotide repeat motifs were found in greater abundance in 5' reads (AAG, AAT, ACC, ACG, ACT, AGG, and CCG) (Fig. 2b), while this was true in four of the eight motifs in spruce (AAC, ACG,

Table 2 Numbers of ESTs containing SSRs before and after assembly in loblolly pine and spruce and divided by direction of the EST reads

Species	Direction	Number of unassembled ESTs with SSRs	After EST assembly		Combining directions	
			Number of contigs with SSRs	Number of singletons with SSRs	Number of contigs with SSRs	Number of singletons with SSRs
Loblolly pine	5'	1,412	138	175	103	334
	3'	418	65	136		
	Unknown	218	13	35		
Spruce	5'	241	38	113	108	491
	3'	956	138	318		
	Unknown	219	33	121		

The results from two EST assemblies are presented; one in which assembly is performed within the read direction classes (5', 3', unknown), and the other in which all reads within species are grouped

Fig. 1 Relative numbers of SSRs found in ESTs of loblolly pine and spruce divided by types. Numbers originate from the recovery of SSRS from a complete assembly of all ESTs within each of the two taxonomic groups



ACT, and AGG). Cumulative weighted trinucleotide frequencies showed no consensus between the two species in terms of EST direction bias. Trinucleotides were 1.6 times more frequent in 5' sequenced ESTs than in 3' sequenced ESTs in loblolly pine, while in spruce, they were 1.2 times more frequent in 3' EST reads than in their 5' counterparts.

Tetranucleotide repeats The two most abundant tetranucleotide motifs in loblolly pine were AAAT (48.0%) and AAAC (12.0%), while in spruce, they were AAGT (26.9%) and ACGG (19.2%) (Table 3). When combining information from both taxonomic groups, only 11 motifs were recovered leaving another 19 possible motifs undetected. The motifs that were found showed a strong base composition bias in which 65.9% of all bases were A+T. Three of the four tetranucleotide repeat motifs listed above as the most frequently encountered in this study were found at greater frequency in 3' sequenced ESTs, while the fourth (AAAC in loblolly pine) was only found in 5' sequenced ESTs (Table 3).

Pentanucleotide and hexanucleotide repeats For pentanucleotide repeats, the two most abundant motifs in loblolly pine were AAAAC (16.7%) and AACTT (16.7%), while in spruce, they were AGAGG (26.6%) and ACCTG (21.1%). For hexanucleotide repeats, the motifs ACGGGG (9.2%) and AAGAGG (7.1%) were the two most abundant in loblolly pine, while in spruce, the most abundant motifs were AGCCGC (8.2%) and ACGGAG and ACCACG (each 4.5%). Twenty-two pentanucleotide repeat motifs were

detected, leaving 44 other possible motifs undetected in this SSR class, but no base composition bias was detected in them. The compilation of recovered hexanucleotide repeats in both species revealed that 94 motifs were recovered, leaving another 130 possible motifs undetected. The close inspection of the recovered motifs also showed no detectable bias in base composition. Cumulative weighted hexanucleotide repeat frequencies showed the same trend as that observed in trinucleotide repeats: the EST direction bias in SSR frequency between loblolly pine and spruce was the opposite.

Compound and imperfect repeats We found a total of 13 compound and 19 imperfect SSRs in this study. Nine of the compound repeats were found in loblolly pine, and 13 of the imperfect repeats were found in spruce. Most compound repeats (6 out of 13) were composed of two dinucleotide motifs followed by those composed of two trinucleotide repeats (5 out of 13). In imperfect repeats, the interrupting sequence varied from 1 to 17 nucleotides in length. The majority of imperfect repeats (10 out of 19) were represented by perfect dinucleotide repeats with a short disruptive element. Upon closer inspection, most imperfect repeats were found to be putatively attributable to sequencing error.

Repeat length distribution

All perfect repeats recovered in this study showed a decrease in their frequency with an increase in their length

Table 3 Number of SSRs found in collections of ESTs from spruce and loblolly pine

Repeat type	Loblolly pine			Spruce		
	5' EST	3' EST	Unknown	5' EST	3' EST	Unknown
(a)						
AC	1	0	1	7	8	5
AG	36	20	2	15	50	17
AT	67	80	9	37	165	60
CG	0	0	0	0	0	0
	104	100	12	59	223	82
(b)						
AAC	5	5	1	3	4	1
AAG	25	10	5	11	26	10
AAT	15	7	3	3	38	12
ACC	13	5	2	3	7	2
ACG	28	15	4	20	24	14
ACT	10	4	2	8	10	3
AGG	20	4	2	17	22	10
CCG	15	5	2	6	14	2
	131	55	21	71	145	54
(c)						
AAAC	3	0	0	0	1	0
AAAG	0	1	0	0	2	0
AAAT	6	5	1	0	2	0
AACT	1	1	0	0	1	0
AAGG	0	0	0	0	1	0
AAGT	0	0	0	0	6	1
AATT	0	1	0	0	0	0
ACCG	1	1	0	0	2	1
ACCT	0	0	0	0	3	0
ACGG	2	0	0	1	3	1
AGAT	1	1	0	0	0	1
	14	10	1	1	21	4
(d)						
Pentanucleotides	37	20	3	16	56	16
Hexanucleotides	84	38	19	26	62	22
Compound	7	2	0	1	3	0
Imperfect	4	2	0	0	13	0

These are categorized by repeat types (e.g., dinucleotides), repeat motifs (e.g., AT), species, and the direction from which the SSR containing ESTs were sequenced (5', 3', or unknown). All numbers were computed from assemblies of loblolly pine and spruce ESTs within read direction and species classes (5', 3', and unknown). Data per motif in penta- and hexanucleotide repeats were collapsed to simplify the table.

(or number of repeat units) as shown in Fig. 3. Di-, tri-, and tetranucleotide repeat frequencies all exhibited a near asymptotic distribution. The point at which the frequency distribution reaches the asymptote appears to coincide with an approximate total SSR length of 30–50 nucleotides (i.e., 15–25 dinucleotide repeats, 10–16 trinucleotide repeats, and 7–12 tetranucleotide repeats).

Discussions

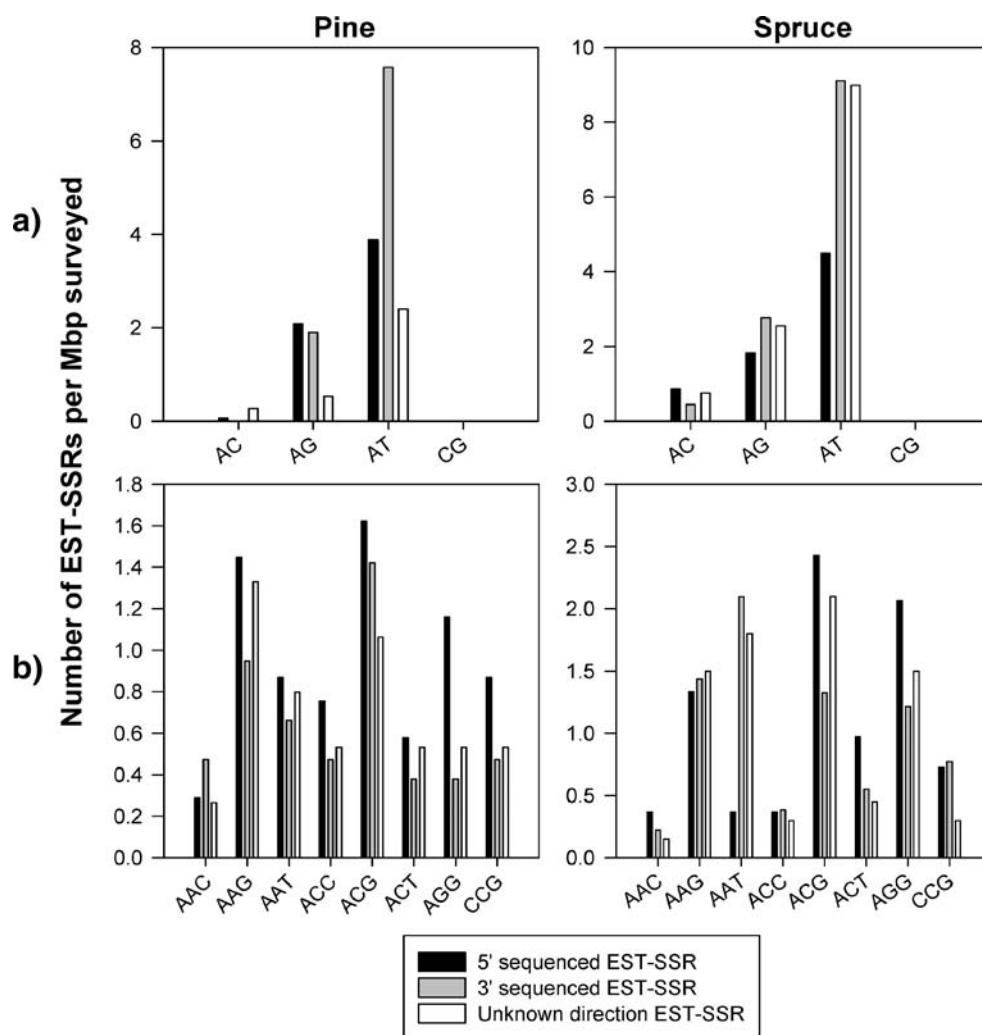
Low frequencies of SSR-ESTs in conifers

The percentage of ESTs containing at least one SSR in this study (1.1% in both loblolly pine and spruce) was lower than that reported by Kantety et al. (2002) for barley, maize, rice, sorghum, and wheat ESTs (1.5 to 4.7%). Average density of EST-SSRs was also low in loblolly pine and

spruce compared to other studies. In loblolly pine and spruce, an SSR occurs in an EST on average every 49.8 kb. In comparison, in EST databases, an EST-SSR occurs every 14 kb in *Arabidopsis* (Cardle et al. 2000), every 19 kb in rice (Temnykh et al. 2001), and every 19.4 kb averaged over wheat, rice, maize, and soybean (Gao et al. 2003). These data suggest that conifers have very low EST-SSR frequency; in fact, they have the lowest of all plants studied to date.

Morgante et al. (2002), in a study of four plant species, reported that whole genome SSR frequency was inversely related to the genome size and to the proportion of repetitive DNA but remained constant over transcribed portions of the genome. Given the large genome size of loblolly pine and spruce and the low SSR recovery rate from ESTs reported in this study, our results suggest that SSR frequency may also decrease in transcribed portions of the genome with an increase in genome size.

Fig. 2 Numbers of di-, tri-, and tetranucleotide repeats per million base pairs surveyed in loblolly pine and spruce ESTs separated by end read type (5' or 3' or unknown). The number of million base pairs surveyed was recovered from assemblies of ESTs



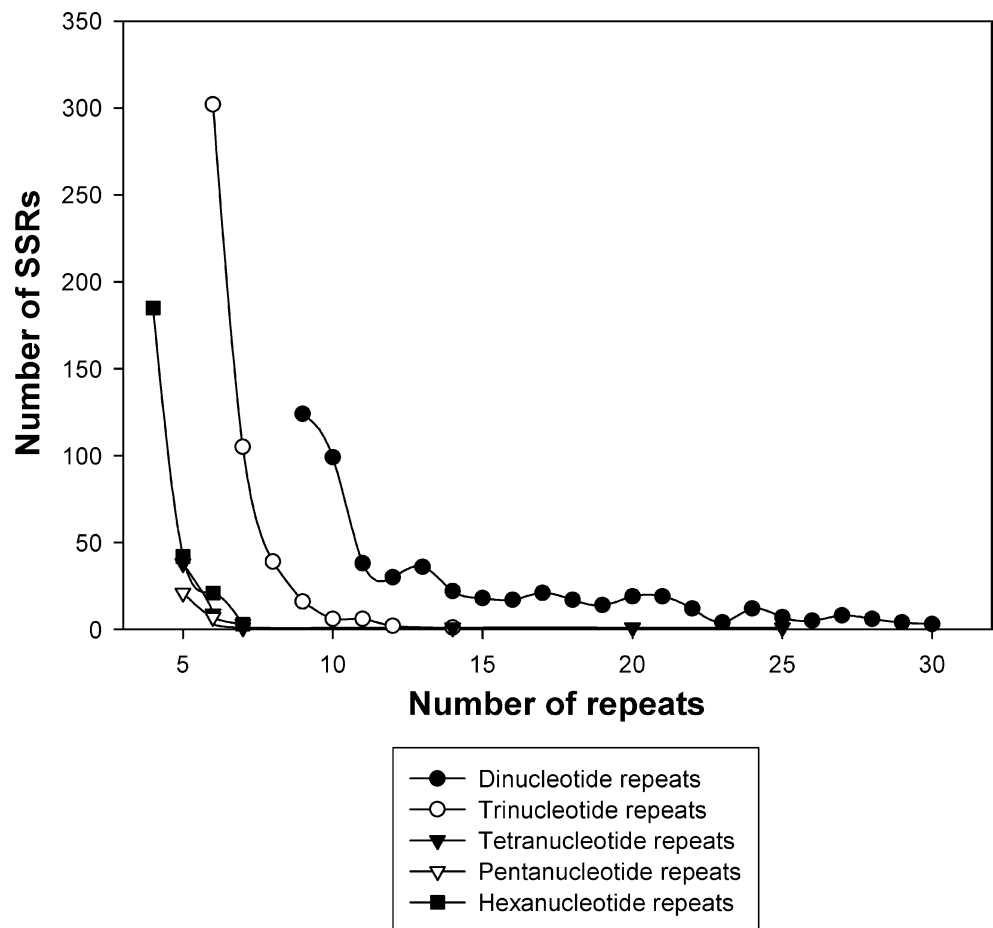
High frequencies of dinucleotide repeats in conifers

Most EST sequences consist of exonic regions, which, because they will be translated into proteins, are under heavy selection against frameshift mutations. As codons are functional units of three nucleotides, indel mutations causing a shift in three nucleotides will not perturb the current reading frame of a given gene (Metzgar et al. 2000). For this reason, trinucleotide repeats are expected to be the most abundant SSR class found in ESTs. Therefore, the observation of dinucleotide repeats as the most abundant SSR class in loblolly pine and spruce ESTs was unexpected and also contrary to the findings of previous studies in which trinucleotide repeats were found to be the most abundant SSR class recovered from plant ESTs (Cardle et al. 2000; Gao et al. 2003; Kantety et al. 2002; Temnykh et al. 2001; and Varshney et al. 2002). The only study proposing a similar trend consists of our preliminary survey of spruce EST data (Rungis et al. 2004). Highly abundant dinucleotide repeats, in fact, are more in line with frequency distribution of SSR classes from genomic DNA (Tóth et al. 2000).

These high frequencies of dinucleotides could be explained as follows. First, dinucleotide repeats in exons could be coding for strings of amino acids that are prevalent in conifers. This explanation has been proposed in cereal species by Kantety et al. (2002) to explain the high abundance of the dinucleotide repeat motif AG observed in their dataset. Alternatively, SSRs are known to be more abundant in introns and intergenic regions than in exons (Hancock 1995): the dinucleotide repeats recovered in this study may have been, in fact, coming from the UTR (untranslated region). Data from our study show that a greater proportion of dinucleotide repeats were recovered from 3' EST reads. Also, 3' UTR lengths have been estimated to be on average 256 bp long based on *in-silico* ORF predictions (a minimum estimate due to many contigs not being complete) in loblolly pine ESTs (Kirst et al. 2003). This may represent a large portion of the sequence composing an EST given that EST length averaged ~500 bp. Hence, it is likely that a large portion of the dinucleotides recovered in this study came from “SSR-rich” 3' UTRs.

As expected, because of potential problems with disturbing the reading frame of exons, trinucleotide and hexanucleotide

Fig. 3 Distribution of the number of repeat units in loblolly pine and spruce ESTs by repeat type



repeats were found to be second and third most abundant classes of SSR, respectively, in both loblolly pine and spruce. If we exclude the findings concerning dinucleotide repeats, the high frequency of trinucleotide repeats found in this study is in accord with previous plant EST-SSR studies (Cardle et al. 2000; Gao et al. 2003; Kantety et al. 2002; Temnykh et al. 2001; Tóth et al. 2000; Varshney et al. 2004). High abundance of hexanucleotide repeats is in agreement with the findings of Gao et al. (2003) in major crops and those of (Tóth et al. 2000) in various eukaryotic genomes but is in contrast with the results of Varshney et al. (2002), where hexanucleotide repeats had the lowest frequency of all surveyed repeats. This is likely a consequence of using a minimum of five repeats as a search criterion compared to two in (Tóth et al. 2000), three in Gao et al. (2003), and four in this study.

Abundance of the AT motif in conifers

Another distinguishing trend observed in conifers was the abundance of the AT dinucleotide repeat motif. This motif was the most frequently recovered one in the dinucleotide class in this study and confirms our observation made on a smaller set of spruce ESTs (Rungis et al. 2004). Also, these results are in

agreement with those of Besnard et al. (2003) where the search for mono-, di-, and trinucleotide repeats in spruce UTRs revealed four out of seven SSRs to be AT motifs.

In contrast, all other plants studied to date exhibit the AG motif as their dominant EST-SSR dinucleotide repeat. This has been reported in *Arabidopsis*, tomato, poplar, and cotton (Cardle et al. 2000), barley (Kantety et al. 2002; Varshney et al. 2002), rice (Cardle et al. 2000; Gao et al. 2003; Kantety et al. 2002; Temnykh et al. 2001; Varshney et al. 2002), tomato (Cardle et al. 2000), maize (Cardle et al. 2000; Gao et al. 2003; Kantety et al. 2002; Varshney et al. 2002), soybean (Cardle et al. 2000; Gao et al. 2003), sorghum (Kantety et al. 2002), wheat (Gao et al. 2003; Kantety et al. 2002; Varshney et al. 2002), and oats and rye (Varshney et al. 2002).

The results presented in this study also differ from those obtained for the conifer Douglas fir (*Pseudotsuga menziesii*). In this species, Southern blots were utilized to evaluate SSR frequency in genomic DNA (Amarasinghe and Carlson 2002). The hybridization of probes representing all 47 possible combinations of di-, tri-, and tetranucleotide repeats revealed the motif AG, AC, and ATG to be the most abundant. Lack of congruence with our results is likely the consequence of surveying the entire genome in Douglas fir as opposed to only the coding regions. Also, Southern blots

depend on hybridization bonding energies; therefore, the relatively weak double hydrogen bond formed between A and T nucleotides are likely to result in an underrepresentation of A+T-rich SSR motifs.

Our results, in fact, are more similar to frequencies of plant SSRs originating from genomic DNA, where the AT motif is most abundant (Morgante and Olivieri 1993; Powell et al. 1996; Temnykh et al. 2001). This adds evidence to the idea that a large number of dinucleotide repeats in spruce and loblolly pine originate from large UTRs found in conifer ESTs. This proposed explanation is in need of further investigation and would be greatly aided by the availability of full-length cDNA sequences, large contiguous genomic sequences, and protein sequence data to determine precisely the position of the UTRs.

Also, the absence or rarity of CG motifs found in this study accords with most other plant dinucleotide EST-SSRs. As demonstrated by (Tóth et al. 2000), this observation also encompasses the intronic, exonic, and intergenic regions of most eukaryotes. In general, CG base pairs are known to be of reduced frequency in many plant and animal genomes.

Other repeat motifs

The frequency distribution of trinucleotide repeat motifs in spruce and loblolly pine was similar to that in wheat, where the most abundant motif was ACG in both coniferous groups (Kantety et al. 2002) and also similar to *Arabidopsis* (Cardle et al. 2000) and soybean (Gao et al. 2003) whose most abundant motif was AAG—the second most abundant trinucleotide motif recovered in loblolly pine. These distributions stand in contrast to those found in monocots, where G+C-rich motifs are typically found to occur at high frequency (Akagi et al. 1996; Gao et al. 2003; Kantety et al. 2002; Temnykh et al. 2001). Moreover, Cardle et al. (2000) proposed that SSR motifs in dicots were typically A+T-rich. In this study, spruce and loblolly pine did not demonstrate any bias in base composition, suggesting that with regard to trinucleotide repeats, the frequency distribution of motifs in conifers is somewhere between dicots and monocots.

Frequency distribution of the large repeat motifs varied greatly between tetra-, penta-, and hexanucleotide repeats. Tetra- and pentanucleotide repeats were underrepresented in the overall SSRs recovered from loblolly pine and spruce. This observation is shared by previous studies: while both types of repeats are relatively frequent in genomic DNA, exonic regions are typically almost devoid of them (Tóth et al. 2000). The base composition bias observed in tetranucleotide repeats of both spruce and loblolly pine, favoring A+T-rich motifs, is similar to the one observed in genomic DNA of all plants surveyed by (Tóth et al. 2000). This similarity, however, should be interpreted cautiously as the sample size for conifers is relatively low. In hexanucleotide

repeats, the most abundant motifs in loblolly pine (ACGGGG, AAGAGG) and spruce (AGCCGC, ACGGAC) suggest a dominance of G+C-rich motifs, but an evaluation of the complete set of recovered hexanucleotide repeats fails to demonstrate a significant bias in base composition. Gao et al. (2003) found that, in monocots, G+C-rich hexanucleotide repeats were more abundant. Similar observations were reported in maize, rice, and vertebrates' cDNA clones (Chin et al. 1996; Temnykh et al. 2001; Tóth et al. 2000). As with base composition of trinucleotide repeats, the G+C-rich genome of monocots is the likely cause for this skewed base distribution among hexanucleotide repeats in this group.

The contribution of compound and imperfect repeats to the overall recovery of SSRs in loblolly pine and spruce was very small. Most studies concerned with *in-silico* recovery of SSRs typically forgo the search for such repeats. It is interesting, however, that most compound repeats recovered in this study were perfect repeats of the same type as opposed to a random assortment of concatenated SSR types. This observation may have implication for the understanding of the functional role or the mutation models of SSRs in general.

SSR length

Across all types of SSRs, those with longer repeat number were less frequent than those with shorter repeat number for both loblolly pine and spruce. This is in agreement with the results from rice (Temnykh et al. 2001) and some cereal species (Varshney et al. 2002). Also, the decrease was more abrupt as repeat units become longer. These results support the hypothesis of Samadi et al. (1998) that loci with longer repeat units experience stronger selection against the difference in size. Because the SSR distribution approaches zero in all repeats when the total SSR length is ~30–50 nucleotides, we hypothesize that this may represent a physical limit to SSR expansion.

Acknowledgments This research was made possible through funding by Genome Canada and the Province of British Columbia (to J.B. and K.R.) through the Genome BC Forestry Genome Project. The authors would also like to acknowledge the support of the Vancouver Genome Sciences Centre for EST sequencing and database development.

References

- Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat Genet* 4:373–380
- Akagi H, Yokozeki Y, Inagaki A, Fujimura T (1996) Microsatellite DNA markers for rice chromosomes. *Theor Appl Genet* 93:1071–1077
- Amarasinghe V, Carlson JE (2002) The development of microsatellite markers for genetic analysis in Douglas-fir. *Can J For Res* 32:1904–1915

- Ayers NM, McClung AM, Larkin PD, Bligh HFJ, Jones CA, Park WD (1997) Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. *Theor Appl Genet* 94:773–781
- Bérubé Y, Ritland C, Ritland K (2003) Isolation, characterization, and cross-species utility of microsatellites in yellow cedar (*Chamaecyparis nootkatensis*). *Genome* 46:353–361
- Besnard G, Acheré V, Faivre Rampant P, Favres JM, Jeandroz S (2003) A set of markers developed from DNA sequence databanks in *Picea* (Pinaceae). *Mol Ecol Notes* 3:380–383
- Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156:847–854
- Chagne D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera MT, Vendramin GG, Garcia V, Frigerio JM, Echt C, Richardson T, Plomion C (2004) Cross-species transferability and mapping of genomic and cDNA SSRs in pines. *Theor Appl Genet* 2004:1204–1214
- Chin ECL, Senior ML, Shu H, Smith JSC (1996) Maize simple repetitive DNA sequences: abundance and allele variation. *Genome* 39:866–873
- Cordeiro G, Casu R, McIntyre C, Manners J, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross-transferable to erianthus and sorghum. *Plant Sci* 160:1115–1123
- Echt CS, May-Marquardt P, Hsieh M, Zahorchak R (1996) Characterization of microsatellite markers in eastern white pine. *Genome* 39:1102–1108
- Gao L, Tang J, Li H, Jia J (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol Breed* 12:245–261
- Gupta PK, Balyan HS, Sharma PC, Ramesh B (1996) Microsatellites in plants: a new class of molecular markers. *Curr Sci* 70:45–54
- Hancock JM (1995) The contribution of slippage-like processes to genome evolution. *J Mol Evol* 41:1038–1047
- Hodgetts RB, Aleksiuk MA, Brown A, Clarke C, Macdonald E, Nadeem S, Khasa D (2001) Development of microsatellite markers for white spruce (*Picea glauca*) and related species. *Theor Appl Genet* 102:1252–1258
- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* 9:868–877
- Jarne P, Lagoda PJJ (1996) Microsatellites, from molecules to populations and back. *Trends Ecol Evol* 11:424–429
- Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. *J Mol Evol* 40:120–126
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum, and wheat. *Plant Mol Biol* 48:501–510
- Khajavi M, Tari AM, Patel NB, Tsuji K, Siwak DR, Meistrich ML, Terry NH, Ashizawa T (2001) ‘Mitotic drive’ of expanded CTG repeats in myotonic dystrophy type1 (DM1). *Hum Mol Genet* 10:855–863
- Kirst M, Johnson A, Baucom C, Ulrich E, Hubbard C, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 100:7383–7388
- Ledig FT, Hodgskiss PD, Krutovskii KV, Neale DB, Eguluz-Piedra T (2004) Relationships among the spruces (*Picea*, Pinaceae) of southwestern North America. *Syst Bot* 29:275–295
- Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. *Theor Appl Genet* 109:361–369
- Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res* 10:72–80
- Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. *Plant J* 3:175–182
- Morgante M, Hanafrey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30:194–200
- Pfeiffer A, Olivieri AM, Morgante M (1997) Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). *Genome* 40:411–419
- Powell W, Machray GC, Provan J (1996) Polymorphisms revealed by simple sequence repeats. *Trends Plant Sci* 1:215–222
- Rajora OP, Rahman MH, Dayanandan S, Mosseler A (2000) Isolation, characterization, inheritance and linkage of microsatellite DNA markers in white spruce (*Picea glauca*) and their usefulness in other spruce species. *Mol Gen Genet* 264:871–882
- Richards RI, Sutherland GR (1992) Dynamic mutations: a new class of mutations causing human disease. *Cell* 70:709–712
- Rungis D, Bérubé Y, Zhuang J, Ralph S, Ritland CE, Ellis BE, Douglas C, Bohlmann J, Ritland K (2004) Robust simple sequence repeat (SSR) markers for spruce (*Picea* spp.) from expressed sequence tags (ESTs). *Theor Appl Genet* 109:1283–1294
- Samadi S, Artigubielle E, Estoup A, Pointier JP, Silvain JF, Heller J, Cariou ML, Jarne P (1998) Density and variability of dinucleotide microsatellites in the parthenogenetic polyploid snail *Melanoides tuberculata*. *Mol Ecol* 7:1233–1236
- Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* 20:211–215
- Scotti I, Egger P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. *Theor Appl Genet* 100:723–726
- Scotti I, Paglia G, Morgante M (2002a) Trinucleotide microsatellites in Norway spruce (*Picea abies* Karst.): their features and the development of molecular markers. *Theor Appl Genet* 106:40–50
- Scotti I, Magni F, Paglia G, Morgante M (2002b) Efficient development of dinucleotide microsatellite markers in Norway spruce (*Picea abies* Karst.) through dot-blot selection. *Theor Appl Genet* 104:1035–1041
- Soranzo N, Provan J, Powell W (1998) Characterization of microsatellite loci in *Pinus sylvestris* L. *Mol Ecol* 7:1260–1261
- Tautz D, Renz M (1984) Simple sequence repeats are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res* 12:4127–4138
- Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* 11:1441–1452
- The Huntington’s Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes. *Cell* 72:971–983
- Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 10:967–981
- van de Ven WTG, McNicol RJ (1996) Microsatellites as DNA markers in Sitka spruce. *Theor Appl Genet* 93:613–617
- Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett* 7:537–546
- Wright JW (1955) Species crossability in spruce in relation to distribution and taxonomy. *For Sci* 1:319–349

Website references

<ftp://ftp.ncbi.nih.gov/repository/dbEST/>; NCBI’s EST database ftp site
<http://www.genetics.forestry.ubc.ca/ritland/programs.html>; Kermit Ritland’s repository of population genetics/bioinformatics programs. The Java program SSRfinder version 2 can be found here