Theor Appl Genet (2009) 119:53–63
DOI 10.1007/s00122-009-1016-3

ORIGINAL PAPER

Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.)

Sylvie Cloutier · Zhixia Niu · Raju Datla · Scott Duguid

Received: 22 January 2009 / Accepted: 14 March 2009 / Published online: 9 April 2009

Abstract A set of 146,611 expressed sequence tags (ESTs) were generated from 10 flax cDNA libraries. After assembly, a total of 11,166 contigs and 11,896 singletons were mined for the presence of putative simple sequence repeats (SSRs) and yielded 806 (3.5%) non-redundant sequences which contained 851 putative SSRs. This is equivalent to one EST-SSR per 16.5 kb of sequence. Trinucleotide motifs were the most abundant (76.9%), followed by dinucleotides (13.9%). Tetra-, penta- and hexanucleotide motifs represented <10% of the SSRs identified. A total of 83 SSR motifs were identified. Motif (TTC/GAA)*n* was the most abundant (10.2%) followed by (CTT/AAG)*n* (8.7%), (TCT/AGA)*n* (8.6%), (CT/AG)*n* (6.7%) and (TC/GA)*n* (5.3%). A total of 662 primer pairs were designed, of which 610 primer pairs yielded amplicons in a set of 23 flax accessions. Polymorphism between the accessions was found for 248 primer pairs which detected a total of 275 EST-SSR loci. Two to seven alleles were detected per marker. The polymorphism information content value for these markers ranged from 0.08 to 0.82 and averaged 0.35. The 635 alleles detected by the 275 polymorphic EST-SSRs were used to study the genetic relationship of 23 flax accessions. Four major clusters and two singletons were observed. Sub-clusters within the main clusters correlated with the pedigree relationships amongst accessions. The EST-SSRs developed herein represent the first large-scale development of SSR markers in flax. They have potential to be used for the development of genetic and physical maps, quantitative trait loci mapping, genetic diversity studies, association mapping and fingerprinting cultivars for example.

Communicated by Y. Xue.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1016-3) contains supplementary material, which is available to authorized users.

S. Cloutier (✉) · Z. Niu
Cereal Research Centre, Agriculture and Agri-Food Canada, 195 Dafoe Road, Winnipeg, MB R3T 2M9, Canada
e-mail: scloutier@agr.gc.ca

R. Datla
National Research Council, Plant Biotechnology Institute, 110 Gymnasium Place, Saskatoon, SK S7N 0W9, Canada

S. Duguid
Morden Research Station, Agriculture and Agri-Food Canada, Route 100, Morden, MB R6M 1Y5, Canada

Introduction

Flax (*Linum usitatissimum* L., $2n = 30$), also called common flax or linseed, is an annual and self-pollinated species. It has been cultivated for its seed oil or stem fibres or both for several 1,000 years (Dillman 1953; Zohary 1999). Traditionally, the oil extracted from the seeds (40–45%) is used for a variety of industrial purposes such as linoleum, paint, varnish, soap and printer ink. The oil-free meal (20–30% protein) is used for feeding livestock. Recently, flax varieties with low linolenic acid have been developed for human consumption (Dribnenki and Green 1995; Dribnenki et al. 1999, 2004). Canada is a major linseed producing country along with Argentina, India, the USA and Russia. Exploitation and characterization of flax genetic resources and evaluation of flax genetic variability are very important for flax germplasm management and breeding.

There are different ways to evaluate flax genetic variation, such as morphological characteristics (Diederichsen

and Hammer 1995; Diederichsen 2001; Diederichsen and Raney 2006; Diederichsen et al. 2006; Saeidi 2008), isozymes (Tyson et al. 1985; Gorman et al. 1993; Krulickova et al. 2002) and molecular markers (Spielmeyer et al. 1998; Oh et al. 2000; Stegnii et al. 2000; Everaert et al. 2001; Fu et al. 2002, 2003a, b; McBreen et al. 2003; Adugna et al. 2005, 2006). Morphological characters tend to be more quantitative and environmentally dependent while isozymes are limited in numbers. DNA markers' abundance, environment insensitivity and non-tissue specific characteristics are some of their advantages. They are useful for variety identification and evaluation of DNA variation. Different molecular markers including random amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP) and simple sequence repeat (SSR) have been developed to analyze flax genetic diversity (Spielmeyer et al. 1998; Oh et al. 2000; Stegnii et al. 2000; Everaert et al. 2001; Fu et al. 2002, 2003a, b; McBreen et al. 2003; Adugna et al. 2005, 2006; Roose Amsaleg et al. 2006). However, their numbers in each of these studies were limited. Some marker types (RAPD and AFLP) have lower cross applicability and others, such as RFLP and AFLP, are quite labour intensive. SSRs, also called microsatellites, consist of a variable number of tandem repeats. They are abundant, evenly distributed, ubiquitous, co-dominant and highly polypmophic (Powell et al. 1996). SSRs can be successfully transferred to related species or different genera (Varshney et al. 2002; Saha et al. 2006; Zhang et al. 2006; Aggarwal et al. 2007). SSR markers can be derived from either genomic sequences or expressed sequence tags (ESTs). EST-based SSRs (EST-SSRs), also called genic SSRs, are particularly informative because they tag the expressed gene from which they were derived and can be used as diagnostic markers for the genes or to map them.

EST-SSRs have been successfully implemented in almond (Xu et al. 2004), Triticeae (Zhang et al. 2006), forage grasses (Mian et al. 2005), coffee (Aggarwal et al. 2007), barley (Varshney et al. 2006), soybean (Hisano et al. 2007), Cucurbita species (Gong et al. 2008), citrus (Chen et al. 2006), potato (Feingold et al. 2005), cotton (Han et al. 2004; Park et al. 2005; Han et al. 2006), rice (La Rota et al. 2005) and wheat (Yu et al. 2004; Peng and Lapitan 2005; Fu et al. 2006). These EST-SSR markers have been used not only for phylogenetic studies but also for the construction of high-density genetic maps. For example, the soybean genetic map, which comprises 935 marker loci, includes 693 EST-SSR loci (Hisano et al. 2007).

SSR markers are still very limited in flax where only 11 and 28 SSR markers have been reported in two independent variety identification studies (Wiesner et al. 2001; Roose Amsaleg et al. 2006). The objectives of the present research were (1) to identify EST-SSRs in flax; (2) to analyse the frequency and distribution of EST-SSRs in the expressed portion of the flax genome; (3) to develop a comprehensive set of novel EST-SSR markers for flax; (4) to assess their polymorphism in a set of 23 flax accessions; and (5) to perform a phylogenetic analysis of 23 flax accessions using a large dataset of EST-SSRs.

Materials and methods

Plant materials and DNA extraction

A set of 23 Canadian flax varieties and breeding lines were grown to the first branching stage in a growth cabinet maintained at 20/16°C with a 16/8 h photoperiod. Accessions and their pedigrees are listed (Table 1). Whole plantlets (100 mg) were collected in liquid nitrogen and lyophilized. DNA was extracted with the DNeasy 96 Plant kit as per manufacturer's instructions (Qiagen, Missisauga, ON), quantified by fluorometry and resuspended at a final concentration of 10 ng/μL.

SSR marker development

Expressed sequence tags were generated from ten cDNA libraries (Table 2). A total of 146,611 ESTs were assembled using CAP3 with the criteria set at 93% identity and 40 bp overlap (Huang and Madan 1999). The assembly resulted in 11,166 contigs and 11,896 singletons for a unigene set of 23,062. Mining for the presence of putative SSRs was performed using the ssr.pl algorithm (Temnykh et al. 2001) and MISA (Thiel et al. 2003). The default criteria for selection of a minimum of nine repeats for dinucleotide motifs, six repeats for trinucleotide motifs and five repeats for tetra-, penta- and hexanucleotide motifs were used. The contig or singleton sequences were used to design primers flanking the putative SSRs. The input criteria for Primer 3.0 (Rozen and Skaletsky 2000) were: a length of 17–23 bp, a GC content of 40–60% and an estimated amplicon size of 200–300 bp. An M13 tail sequence (5′-CACGACGTTGTAAAACGAC-3′) was added to all forward primers.

SSRs detection

A total of 50 ng of genomic DNA from each of the 23 flax accessions (Table 1) was used as template for PCR reactions which were performed in 384 well plates with a final volume of 10 μL per reaction. PCR reactions and conditions were previously described (Huang et al. 2006). A total of 2 μL each of FAM-labelled, HEX-labelled and NED-labelled reactions, representing three different putative SSRs, were

Table 1 List of accessions used in assessment of EST-SSRs and their pedigrees

Accession ID	Pedigree/source	References
AC Carnduff	Dufferin//NorLin/Culbert 79	Kenaschuk and Rashid (1999)
AC Emerson	Noralta/Vimy (Vimy = Linott/Kubanskij)	Kenaschuk et al. (1996)
AC Linora	NorMan/Linott	Kenaschuk and Rashid (1993)
AC McDuff	FP766/FP775 (FP766 = Dufferin//Redwood 65/Linott FP775 = McGregor//Redwood 65/HIGH OIL LINE)	Kenaschuk and Rashid (1994)
AC Watson	NorLin//FP775/CI2941	Kenaschuk and Rashid (1998)
CDC Bethune	NorMan/FP857 (FP857 = FP705/Norlin)	Rowland et al. (2002)
Hanley	Flanders/AC Emerson	Duguid et al. (2003a, b, c)
Lirina	Unknown	
MacBeth	M2701/AC Linora	Duguid et al. (2003a, b, c)
Prairie blue	FP956/Flanders	Duguid et al. (2004)
Double low	Unknown	
FP2102	M3286/AC Emerson	Duguid (personal communication)
FP2107	AC Linora/M2348	Duguid (personal communication)
FP2137 (Prairie Thunder)	FP974/FP1043 (FP974 = AC Watson; FP1043 = AC Emerson/AC Linora)	Duguid (personal communication)
FP2159	FP1043/AC Watson	Duguid (personal communication)
FP2161 (Prairie Grande)	AC Watson/CI3395	Duguid (personal communication)
M5791	95-27018-3 (F5)/95-27021-4 (F5) [95-27018-3 (F5) = 93-14492(B)/95-15117(Y); 95-27021-4 (F5) = 95-15117Y/92-235-4B)]	PRH report (2006)
M6552	97-7981-1/UGG5-5 (97-7981-1 = 95-27018-3/97-27021-4) (95-27018-3 = 93-14492B/93-15117Y; 97-27021-4 = 93-15117Y/92-235-4B)	
SP2126	SP 992/94-7889 (SP 992 = NorMan/Zero//CPI 84495/3/ NorMan, 94-7889 = 90-5889/AC McDuff)	Dribnenki et al. (2005)
SP2148	SP 2013/96-129-F2	Dribnenki (personal communication)
SP2149	1084/96-32-F3 (96-32-F3 = 989/95-1002)	Dribnenki et al. (2007)
SP2047	989//Windermere/M2702 (989 = McGregor/Zero//CPI 84495/3/McGregor)	Dribnenki et al. (2003)
UGG5-5	Unknown	

Table 2 Description of the ten cDNA libraries used for the generation of flax ESTs

Accession	Tissue and stage	Number of ESTs	Source	Origin
AC McDuff	Bolls 12 days after flowering	5,093	GenBank	S. Cloutier, Canada
Hermes	Stem outer fiber	927	GenBank	G. Neutelings, France
CDC Bethune	Globular embryos	15,989	NAPGEN[a]	R. Datla, Canada
CDC Bethune	Heart stage embryos	17,185	NAPGEN	R. Datla, Canada
CDC Bethune	Torpedo stage embryos	16,926	NAPGEN	R. Datla, Canada
CDC Bethune	Bent stage embryos	20,213	NAPGEN	R. Datla, Canada
CDC Bethune	Mature stage embryos	15,771	NAPGEN	R. Datla, Canada
CDC Bethune	Pooled endosperm	17,112	NAPGEN	R. Datla, Canada
CDC Bethune	Endosperm at globular stage	20,967	NAPGEN	R. Datla, Canada
CDC Bethune	Seed coat at torpedo stage	16,428	NAPGEN	R. Datla, Canada
Total		146,611		

[a] Natural Product Genomic Consortium

combined with 24 μL of water. An aliquot of 2 μL of the combined products were mixed with 3.9 μL of Hi-Di formamide and 0.1 μL of Genescan ROX-500 standard (Applied Biosystems, Foster City, CA), denatured 5 min at 95°C and transferred onto ice for 5 min before being resolved on an ABI 3100 or 3130xl DNA analyzer (Applied

Biosystems). Output files were analyzed by GeneScan (Applied Biosystems) and subsequently imported into Genographer or, alternatively, the ".fsa" files were directly imported into Genographer (Benham et al. 1999) as modified by T. Banks for the SSR data resolved on ABI DNA analyzers (http://sourceforge.net/projects/genographer). The three labelled reactions were transformed into independent gel-like images. Fragment sizes were estimated using the GeneScan ROX-500 internal size standard and recorded for each accession. Amplicons larger than 450 bp were redone using the MapMarker® 1000 (BioVentures Inc., Murfreesboro, TN) internal size standard.

Polymorphism information content

The polymorphism information content (PIC) is a measure of the effectiveness of a given DNA marker for detecting polymorphism. The PIC value for each EST-SSR marker was calculated by using the standard formula (Anderson et al. 1993):

$$\mathrm{PIC}_i = 1 - \sum_{j=1}^{k} P_{ij}^2.$$

here, P_{ij} is the frequency of the jth allele for the ith marker in a set of the investigated flax accessions and summation extends over k alleles detected for the ith marker.

Phylogenetic analysis

The EST-SSR alleles were converted into a binary matrix where the presence of an allele was scored as 1, its absence was scored as 0 and missing data were given the value −999 as per NTSYS data format (Exeter Software, Setauket, NY). The resample function was used to create 9,999 randomization of the data set and used to obtain the bootstrap values. Genetic similarities were calculated among all possible pairs of accessions using the Dice coefficient and the corresponding UPGMA (Unweighted Pair-Group Method Arithmetic Average) phenograms were generated with the NTSYSpc (v2.2) program (Exeter Software, Setauket, NY).

Results

Frequency and distribution of SSRs in the ESTs of flax

A total of 146,611 EST sequences from 10 cDNA libraries were assembled into a 23,062 unigene set representing 11,166 contigs and 11,896 singletons (Fig. 1). This assembly totalled ~14 Mb thereby representing approximately 2% of the estimated 700 Mb flax genome (Benneth and Leitch 2005). A total of 806 EST sequences representing

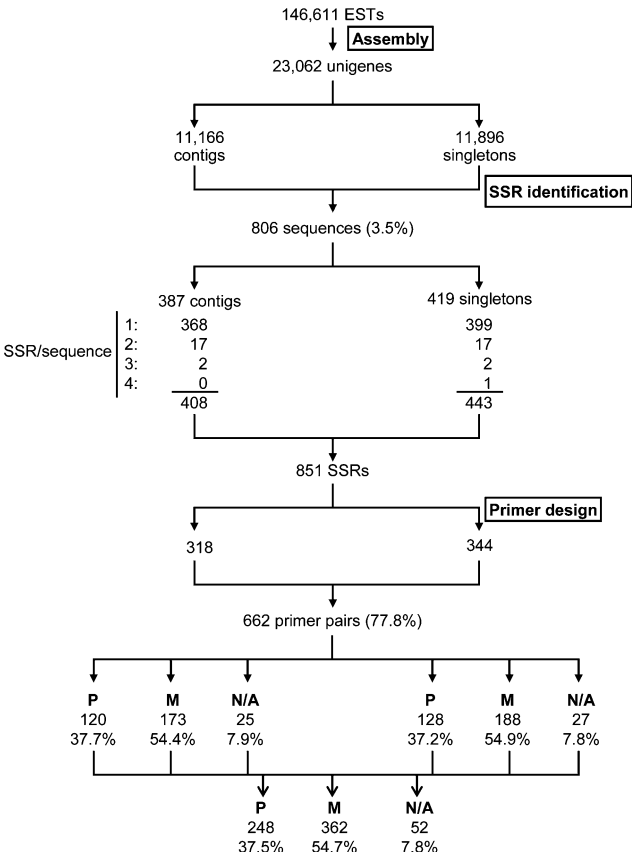

Fig. 1 Diagram illustrating the EST assembly, the SSR identification, the primer design and the polymorphism of the flax EST-SSRs. *P* polymorphic, *M* monomorphic, *N/A* no amplification

0.517 Mb of DNA sequences comprised putative SSRs. The majority (767/806, 95.2%) of the sequences had a single putative SSR, while 39 (4.8%) of them had 2, 3 or 4 putative SSRs per sequence (Fig. 1) for a total of 851 putative SSRs. Therefore, 3.5% of the unigene sequences had at least one SSR. The frequency of occurrence for EST-SSRs averaged one SSR per 16.5 kb of EST sequence. A total of 662 primers were designed. Primers could not be designed for the remaining putative EST-SSRs because (1) the tandem repeats were too close to either end of the sequence, (2) the nature of the sequence did not allow for primer design using Primer 3.0 selection criteria or (3) primers were designed for only one of the putative SSRs of a sequence containing multiple targets. A list of the primer sequences can be found in Table S1 (Supplementary data).

Polymorphism analysis of EST-SSRs markers

All 662 primer pairs were tested on the set of 23 flax accessions (Table 1). A total of 248 primer pairs (37.5%) were polymorphic, 362 primer pairs (54.7%) were monomorphic and 52 (7.8%) failed to amplify any of the 23 genomic DNA templates even after reducing the annealing

temperature by 8°C (Fig. 1). The majority of the primer pairs amplified a single polymorphic locus but a small number of them amplified two or three polymorphic loci. Overall, the 248 polymorphic primer pairs detected 275 loci, which were used to fingerprint the 23 flax accessions.

The majority of the markers were bi-allelic but the number of alleles per EST-SSR varied from 2 to 7 (Supplementary data Table S2). EST-SSR marker Lu685, shown in Fig. 2, had 7 alleles ranging from 391 to 437 bp. A total of 635 alleles were generated by the 275 EST-SSRs for an average of 2.31 alleles per marker. PIC values of the 275 EST-SSRs ranged from 0.08 to 0.82 and averaged 0.35 (Supplemental data Table S2).

Distribution of flax EST-SSRs based on the motif sequence

A total of 83 different SSR motifs were identified in the unigene set. The top 29 motifs (any two complementary sequences were considered one motif) are listed in Table 3. They represent 84.6% of the putative EST-SSR sequences while the remaining 54 motifs accounted for only 15.4% of the 851 EST-SSRs detected. Trinucleotide motifs TTC/GAA, CTT/AAG and TCT/AGA were the most abundant with frequencies of 10.2, 8.7 and 8.6%, respectively. They were followed by dinucleotide motifs CT/AG and TC/GA which had frequencies of 6.7 and 5.3%, respectively.

Distribution of flax EST-SSRs based on their motif, number of repeats and length of the SSR

Frequency distribution based on motifs is illustrated in Fig. 3a. Trinucleotide motifs were the most common repeat, with a frequency of 76.9%, followed by di- (13.9%), tetra- (6.8%), penta- (1.4%) and hexanucleotide (1.0%) motifs.

The number of motif repeats varied from 5 to 21 (Fig. 3b). Six was the most frequent repeat number. An inverse relation between the frequency of putative EST-SSRs was found with increasing number of repeats. The

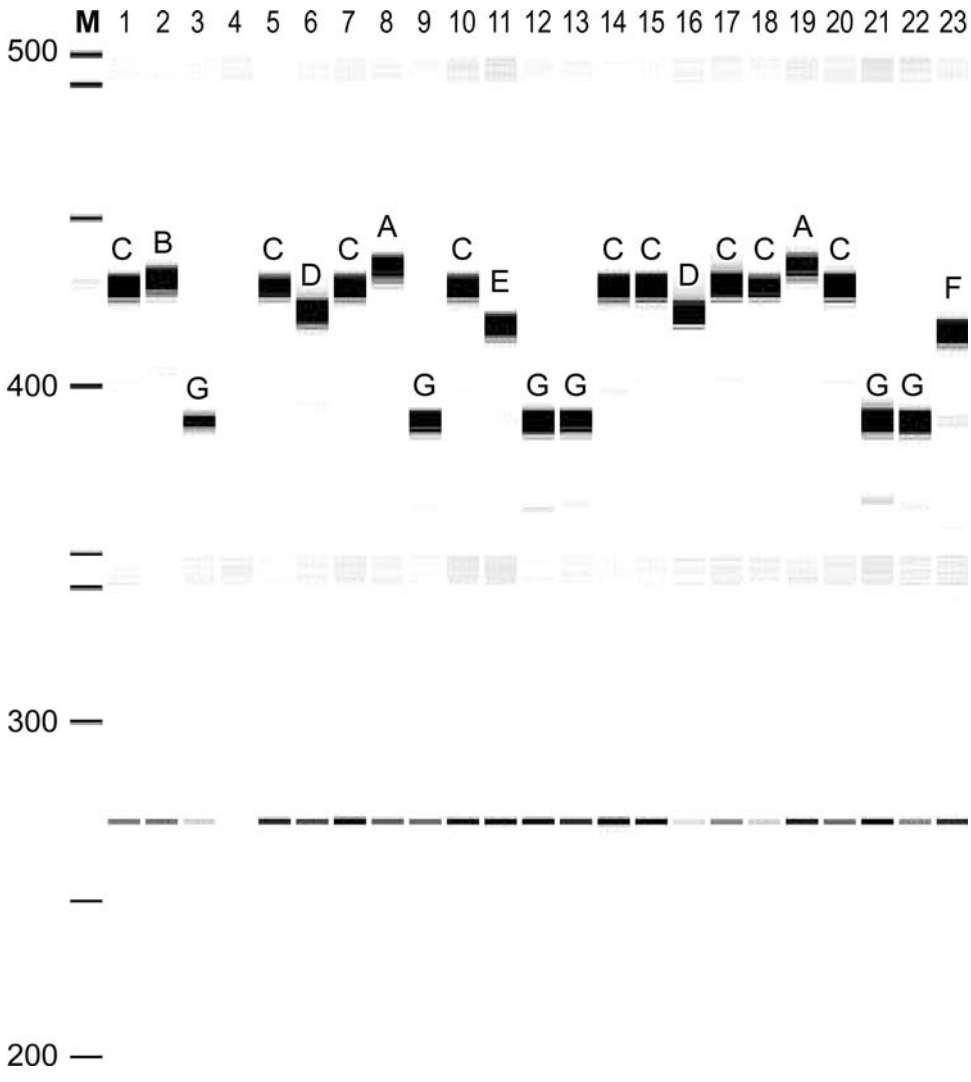

Fig. 2 Gel-like image of EST-SSR Lu685 showing the presence of 7 alleles for this marker amongst the 23 accessions. Alleles are labelled *A–G* and represent estimated sizes of 437, 433, 431, 423, 419, 417 and 391 nucleotides, respectively. Numbers *1–23* at the top refer to the accessions as listed in Table 1. Molecular sizes are indicated on the *left*. *M* marker was GeneScan Rox-500 (Applied Biosystems)

Table 3 Occurrence and number of repeats of the most frequent EST-SSR motifs in flax

Repeat motifs	Number of repeat units											Total
	5	6	7	8	9	10	11	12	13	14	15+	
TTC/GAA		38	20	13	8	5	0	0	0	1	2	87
CTT/AAG		36	15	11	4	3	3	0	0	0	2	74
TCT/AGA		29	21	6	5	5	1	2	1	0	3	73
CT/AG		0	0	0	9	10	12	10	4	4	8	57
TC/GA		0	0	0	10	9	8	7	3	1	7	45
GCT/AGC		24	9	2	1	1	0	0	0	0	0	37
TCA/TGA		19	8	5	0	1	0	0	0	0	0	33
CTG/CAG		18	9	1	1	0	1	0	0	0	0	30
TCC/GGA		10	15	1	1	0	0	0	0	0	0	27
CAT/ATG		13	6	5	0	0	0	0	0	0	0	24
ACA/TGT		9	6	2	1	1	0	0	0	0	0	19
TGC/GCA		13	4	2	0	0	0	0	0	0	0	19
TTA/TAA		10	6	2	0	0	0	0	0	1	0	19
TAT/ATA		11	3	0	2	0	1	0	0	1	0	18
AT/AT					9	5	2				1	17
ATC/GAT		10	4									14
CAA/TTG		6	5	3								14
GAT/ATC		7	4	3								14
TA/TA					8	1	4		1			14
AAT/ATT		5	3	3	1		1	1				14
CTC/GAG		9	3	2	0	0	0	0	0	0	0	14
ATT/AAT		3	3	1	3							10
TTG/CAA		5	3		1		1					10
AGT/ACT		4	1		2	1					1	9
ACC/GGT		6	1		1							8
AAC/GTT		2		1	2							5
CCT/AGG		5										5
GTT/AAC		3	2									5
TTTC/GAAA	4	1										5
Other motifs	43	55	15	6	10		1	1				131
Total	47	351	166	69	79	42	35	21	9	8	24	851

motifs repeating more than 12 times had very low frequency of less than 1%. There were also fewer putative EST-SSRs with five repeat units but this determination was biased because only tetranucleotide motifs and higher were considered with five repeats.

Figure 3c shows the distribution pattern of EST-SSRs based on the length of SSRs (motif size × number of repeats). The highest frequency was found for SSRs that were 18 (44.5%), 21 (18.6%) and 24 (11.8%) nucleotides in length. Frequency decreased dramatically for SSR lengths of 25–60 nucleotides.

SSR gene disruption

Tri- and hexanucleotide SSRs do not cause frame shifts when present in ESTs because they are multiples of three, i.e., the number of nucleotides in a codon. This is not the case for di-, tetra- and pentanucleotide motifs. To estimate the number of EST-SSRs that could potentially result in disruption of the gene where they reside, we evaluated the alternate alleles produced by each EST-SSR. These results are summarized in Table 4. To derive this information, we assumed that the allele size from the genotype from which the primers were designed was a no frame shift allele (NFSA) because it was expressed. The other allele sizes were determined and alternate alleles that would not engender a frame shift were determined. For example, Lu505 was designed from a CDC Bethune EST sequence and amplified a 307 nucleotide fragment in this variety. The SSR was identified as nine repeats of the dinucleotide motif TC/GA. Lu505 detected only 2 alleles and the alternate allele was 301 nucleotides in length representing

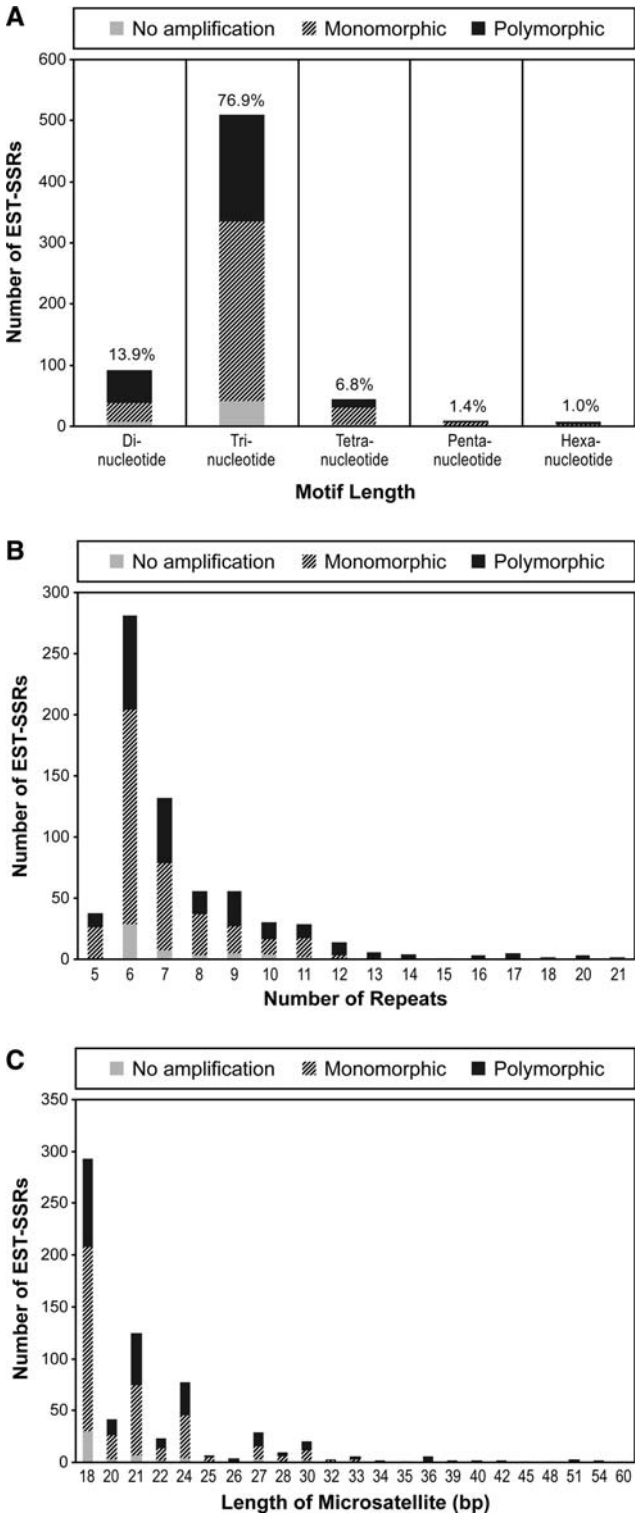

Fig. 3 Frequency distribution of the putative EST-SSRs from flax ESTs based on **a** motif length, **b** number of repeats and **c** length of microsatellite (motif length × number of repeats)

6 repeats of the TC motif. This alternate allele would not cause a frame shift even if it was exclusively in the coding sequence. Alternatively, SSRs located in 5′ or 3′ UTRs would not be disruptive regardless of the SSR motif. To estimate the occurrence of such SSRs, we performed a BLAST search against the Arabidopsis CDS sequence (TAIR build 8_20080412). Twenty-four sequences containing a di-, tetra- or pentanucleotide motif matched *Arabidopsis* ORFs. In 20 cases, the SSRs were located in the 3′ UTR and in the other 4 cases, the SSRs were in the 5′ region. EST-SSRs may not be widely disruptive providing that the di-, tetra- and pentanucleotide motifs be located in the non-coding sequences.

Phylogenetic analysis

Allelic data from the 275 polymorphic EST-SSR markers were used to study the genetic diversity and genetic relationship of 23 Canadian flax accessions. A total of 635 alleles were obtained in all 23 flax accessions.

The dendogram representing the cluster analysis of the 23 flax accessions is shown in Fig. 4. A total of four main clusters comprising 3–8 genotypes each can be observed. Variety Lirina and accession Double Low did not cluster with any other accessions. The two main clusters at the top of Fig. 4 could be further divided into three and two subclusters, respectively. The SP lines all clustered together. Similarly, high linolenic acid accessions M5791, M6552 and UGG5-5 formed their own cluster. FP2137, FP2159 and FP2161 clustered with cultivar AC Watson which is present in their pedigree (Table 1). This pedigree correlated clustering was generally the case. A notable exception was FP2107 (AC Linora/M2348) which had a 0.91 coefficient of similarity with FP2102 but did not cluster with AC Linora despite having the same allele at 66% of the loci (data not shown).

Discussion

EST-SSRs frequency and distribution

Simple sequence repeats were detected in approximately 3.5% of the unigene set obtained from the assembly of more than 146,000 flax ESTs, which is higher than the previous reports in grapes (2.5%) (Scott et al. 2000) and barley (2.8%) (Varshney et al. 2006) but lower than coffee (18.5%) (Aggarwal et al. 2007) and wheat (7.41%) (Peng and Lapitan 2005). The average distance (in kb) between EST-SSRs was 3.4 in rice, 5.46 in wheat, 6.3 in barley, 7.4 in soybean, 8.1 in maize, 11.1 in tomato, 13.8 in *Arabidopsis*, 14 in poplar and 20 in cotton (Cardle et al. 2000; Thiel et al. 2003; Peng and Lapitan 2005). The 16.5 kb interval found herein for flax EST-SSRs is similar to the larger interval reported for cotton indicating that they are less prevalent in flax than in other important and model

Table 4 Characterization of EST-SSRs based on their motifs and the frequency of alleles potentially causing frameshifts that could disrupt the expression of the resident gene(s)

Motif	Primers amplified	Primers polymorphic	% polymorphic	Loci detected	Alleles detected	Allele:loci ratio	No Frame Shift Alleles (NFSA)	% NFA
Di	85	53	62.4	55	150	2.73	64	42.7
Tri	467	174	37.3	196	435	2.22	429	100
Tetra	42	14	33.3	17	34	2	17	50
Penta	9	3	33.3	3	6	2	4	66.7
Hexa	7	4	57.1	4	10	2.5	10	100
Total	610	248	40.7	275	635	2.31	524	82.5

plant species. However, these numbers refer to the identification of putative EST-SSRs using various SSR mining software and does not reflect the level of polymorphism of these SSRs in their respective species. Moreover, the large difference in the average distance between EST-SSRs could also have resulted from the use of different SSR search criteria, the size of databases and the database mining tool(s) used (Varshney et al. 2005).

Trinucleotide motifs were the most abundant, i.e., 76.8%, which is five times higher than the next abundant motif size (dinucleotide 13.9%). This was in agreement with EST-SSR distribution reported in barley, grape, rice, wheat, citrus, cotton and soybean (Scott et al. 2000; Thiel et al. 2003; Han et al. 2004; La Rota et al. 2005; Chen et al. 2006; Hisano et al. 2007) and may simply be reflecting the fact that trinucleotide motifs in coding regions would not cause frame shift mutation that could silence the gene but would result in a variation in amino acid residue number at the protein level. However, in peach, pumpkin, coffee, spruce and kiwi fruit, dinucleotide motif EST-SSRs were the most abundant (Fraser et al. 2004; Rungis et al. 2004; Xu et al. 2004; Aggarwal et al. 2007; Gong et al. 2008). The relative abundance of di- and trinucleotide motifs detected is also a function of the SSR searching criteria and the software used for EST database mining and could have partly contributed to the observed differences (Varshney et al. 2005; Aggarwal et al. 2007).

The most common SSR motif was the trinucleotide and the most common SSR lengths (Fig. 3c) were 18, 21 and 24, all multiple of three, which support the idea that there is a positive selection pressure for in-frame SSRs in genic sequences. Assuming that all EST-SSRs were located in coding sequences only, 44.7% of the alleles detected by dinucleotides would not cause a frame shift. This surprisingly low frequency would indicate that the remaining SSR alleles could prevent the expression of the gene where they reside unless the primers amplified an intron or portion of the 5′- and 3′-untranslated regions (UTRs). Too few tetra- and pentanucleotide EST-SSRs were assessed to draw a strong conclusion for these motifs. However, it is interesting to note that all 17 loci detected by tetranucleotide motifs had only two alleles per locus and that all alternate alleles would cause a frame shift, again assuming their location in coding sequences.

Abundance of CCG/CGG motifs was reported to be a specific feature of monocot genomes (Yu et al. 2004; Peng and Lapitan 2005). GC-rich SSR motifs were found to be less frequent in ESTs from dicots, where GA/TC, AT/AT and/or AG/CT were the most frequent dinucleotide motifs and, CTT/AAG and/or TTC/GAA were the most frequent trinucleotide motifs (Cardle et al. 2000; Scott et al. 2000; Morgante et al. 2002; Rungis et al. 2004; Xu et al. 2004; Feingold et al. 2005; Saha et al. 2006; Gong et al. 2008). These results were corroborated in flax where trinucleotide motifs TTC/GAA (10.2%) and CTT/AAG (8.7%) and dinucleotide motifs CT/AG (6.7%) and TC/GA (5.3%) were the most frequent.

Marker development and polymorphism of EST-SSR markers

In comparison with the reports in sugarcane and barley, where approximately 40% and 30–36% of the designed EST-SSR primers failed, respectively (Cordeiro et al. 2001; Thiel et al. 2003; Varshney et al. 2006), only 7.8% of the primers designed from flax ESTs failed to amplify the 23 DNA templates. These failures could be caused by primer mismatch, primers designed across splice sites or the presence of large introns within the target amplicon. The small portion of failed primers probably also reflect the high quality of the flax EST sequence data.

A total of 248 of the 610 primer pairs that amplified (40.7%) were polymorphic for a set of 23 flax accessions comprised mostly of Canadian varieties and breeding lines. This is higher than other plant species such as wheat (25%), barley (35% and 42%), soybean (12.8%) and cotton (18.2%) (Eujayl et al. 2002; Thiel et al. 2003; Han et al. 2004; Varshney et al. 2006; Hisano et al. 2007). Differences in the polymorphism rates might be partly attributable to the relatedness or the number of genotypes

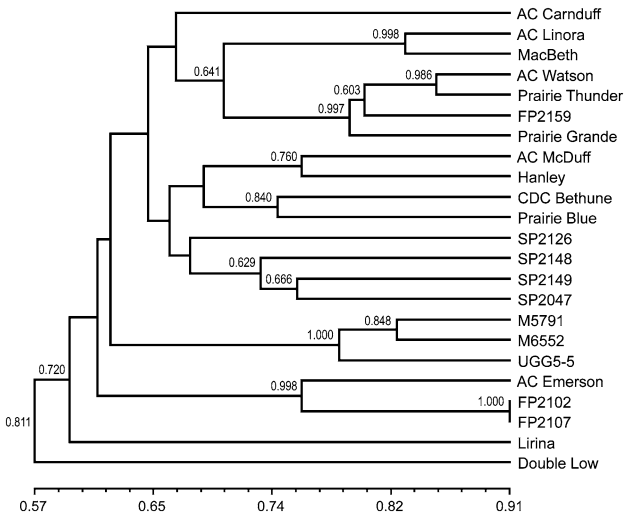

Fig. 4 Dendrogram of the cluster analysis of 23 flax accessions based on 275 EST-SSR markers analyzed by the UPGMA clustering method. Bootstrap values greater than 0.5 are shown

tested. Ploidy level may also affect the proportion of polymorphic EST-SSRs. Indeed, disrupted EST-SSRs would likely be under less selection pressure in a polyploid plant because the homoeologues could compensate for the non-expression of one of the copies. Generally, EST-SSRs show lower levels of polymorphism than genomic SSRs (Eujayl et al. 2002). Roose Amsaleg et al. (2006) reported that 28 genomic SSR markers exhibited 2–10 alleles per locus with an average of 3.32 when assessed on 93 flax cultivars. We observed 2–7 alleles per locus with an average of 2.3 based on 248 EST-SSRs assessed on 23 flax accessions. The PIC value difference between the two studies may be a reflection of the nature of the SSRs (genomic versus EST-derived), the relatedness of the genotypes or simply the number of genotypes tested. When compared to EST-SSR polymorphism from other species, flax was similar to soybean (2.8) and some grass species (1.6–2.5) (Mian et al. 2005; Hisano et al. 2007) but lower than almond (5.45) (Xu et al. 2004). There was, however, a wide variation in PIC value and while the majority of the EST-SSRs had only two alleles, four EST-SSRs (Lu273, Lu685, Lu628 and Lu236) showed high level of polymorphism (PIC $>$ 0.70) indicating that highly polymorphic SSR markers could be obtained from flax ESTs (Supplementary data Table S2).

Dinucleotide motif EST-SSRs showed polymorphism more frequently than trinucleotide motif EST-SSRs (La Rota et al. 2005; Hisano et al. 2007) as was the case in flax where 57.6% of the dinucleotide motifs were polymorphic as compared to 34.2% for the trinucleotide motifs (Fig. 3a). Also, the average PIC value of the dinucleotide motif EST-SSRs (0.37) was higher than that of the trinucleotide motif EST-SSRs (0.33) with TC/GA and CT/AG motif EST-SSRs with the highest overall values of any given motif (Supplementary data S2).

Genetic relatedness of 23 flax accessions

The 23 flax accessions used in this study are a sample of current Canadian breeding material. They were not selected for extent of genetic variability. On the contrary, they form a rather narrow genetic base as compared to the potential genetic variability that can be found in *L. usitatissimum* (L.). The main goal was to develop markers useful in breeding. The clustering analysis shown in Fig. 4 generally corroborated the pedigree information of the lines included in the study (Table 1). The clustering of all high linolenic acid lines (UGG5-5, M5791 and M6552) is a good example because it reflects their common genetic background (95-27018-3). The EST-SSR markers identified in this study successfully disclosed useful genetic variability among flax accessions and could be important resources for researchers working with flax and related species. Designing crosses for specific breeding purposes or to maximize genetic variability would be aided by the EST-SSR information. The EST-SSRs could also be applied in fingerprinting for variety identification (Wiesner et al. 2001; Roose Amsaleg et al. 2006), in association mapping studies or in designing mapping populations. The average similarity of the 23 accessions was 0.64 (0.48–0.91). AC Emerson and Double Low only had 48% similarity (data not shown, similarity matrix). Such a cross could be very useful in mapping EST-SSRs because 137/255 markers (54%) assessed for both accessions could be mapped with this single cross.

Cross-species transferability and application of flax EST-SSRs

EST-SSRs are codominant and multi-allelic. They represent putative functional sequences and were found to be transferable to related species due to the high level of conservation in the flanking sequences of SSRs (Scott et al. 2000; Feingold et al. 2005). Searches for novel alleles in a broader gene pool of related *Linum* species or simply a genetic collection of *L. usitatissimum* (L.) including landraces and world collections, will be facilitated and enhanced by EST-SSRs. Xu et al. (2004) found that some EST-SSRs that showed no polymorphism in common almond were highly polymorphic in other *Prunus* species. The 362 monomorphic primer pairs assessed in this study might display polymorphism in a broader flax collection or in other species of the genus *Linum*.

A large number of *Linum* species have been described from the world *ex situ* collections (Diederichsen 2007).

Inter-specific crossing ability has only been demonstrated for a limited number of species, all of which also had 30 chromosomes (see review in Diederichsen 2007). Pale flax [*Linum bienne* (Mills), *Linum usitatissimum* L. subsp. *angustifolium* (Huds.)] belongs to the primary gene pool of common flax and has the most potential for novel gene/allele transfer to *L. usitatissimum* (L.). EST-SSRs may unravel novel alleles from these species that could be transferred to common flax by inter-specific hybridization.

This report is the first comprehensive study on the development and analysis of a large set of SSR markers in flax. The markers have applications in breeding and will be used for basic and applied research activities.

Acknowledgments The authors are grateful to Andrzej Walichnowski for manuscript review, Joanne Schiavoni for manuscript preparation and Mike Shillinglaw for the preparation of figures. Technical assistance of Laura Marginet, Natalie Middlestead, Mira Popovic, Elsa Reimer and Andrzej Walichnowski is also acknowledged. Travis Banks provided most helpful bioinformatics expertise throughout the project. This work was financially supported by AAFC A-base project for high-value flax development. This is AAFC contribution number 1972.

References

Adugna W, Viljoen CD, Labuschagne MT (2005) Analysis of genetic diversity in linseed using AFLP markers. Sinet Ethiop J Sci 28:41–50

Adugna W, Labuschagne MT, Viljoen CD (2006) The use of morphological and AFLP markers in diversity analysis of linseed. Biodivers Conserv 15:3193–3205

Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor Appl Genet 114:359–372

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Benham J, Jeung J-U, Jasieniuk M, Kanazin V, Blake T (1999) Genographer: a graphical tool for automated fluorescent AFLP and microsatellite analysis. J Agric Genom 4. http://wheat.pw.usda.gov/jag/

Benneth MD, Leitch IJ (2005) Nuclear DNA amounts in angiosperms—progress, problems and prospects. Ann Bot 95:45–90

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

Chen C, Zhou P, Choi YA, Huang S, Gmitter FG Jr (2006) Mining and characterizing microsatellites from citrus ESTs. Theor Appl Genet 112:1248–1257

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160:1115–1123

Diederichsen A (2001) Comparison of genetic diversity of flax (*Linum usitatissimum* L.) between Canadian cultivars and a world collection. Plant Breed 120:360–362

Diederichsen A (2007) Ex situ collections of cultivated flax (*Linum usitatissimum* L.) and other species of the genus *Linum* L. Genet Resour Crop Evol 54:661–678

Diederichsen A, Hammer K (1995) Variation of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum*) and its wild progenitor pale flax (subsp. *angustifolium* (Huds.) Thell.). Genet Res Crop Evol 42:263–272

Diederichsen A, Raney JP (2006) Seed colour, seed weight and seed oil content in *Linum usitatissimum* accessions held by Plant Gene Resources of Canada. Plant Breed 125:372–377

Diederichsen A, Raney JP, Duguid SD (2006) Variation of mucilage in flax seed and its relationship with other seed characters. Crop Sci 46:365–371

Dillman AC (1953) Classification of flax varieties, 1946. US Dept. of Agriculture, 1953. SERIES INFORMATION: Technical bulletin/United States Department of Agriculture; no. 1064. PUBLISHER: US Dept of Agriculture, Washington

Dribnenki JCP, Green AG (1995) Linola '947' low linolenic acid flax. Can J Plant Sci 75:201–202

Dribnenki JCP, McEachern SF, Green AG, Kenaschuk EO, Rashid KY (1999) Linola '1084' low-linolenic acid flax. Can J Plant Sci 79:607–609

Dribnenki JCP, McEachern SF, Chen Y, Green AG, Rashid KY (2003) LinolaTM 2047 low linolenic acid flax. Can J Plant Sci 83:81–83

Dribnenki JCP, McEachern SF, Chen Y, Green AG, Rashid KY (2004) 2090 low linolenic acid flax. Can J Plant Sci 84:797–799

Dribnenki JCP, McEachern SF, Chen Y, Green AG, Rashid KY (2005) 2126 low linolenic flax. Can J Plant Sci 85:155–157

Dribnenki JCP, McEachern SF, Chen Y, Green AG, Rashid KY (2007) 2149 solin (low linolenic flax). Can J Plant Sci 87:297–299

Duguid SD, Kenaschuk EO, Rashid KY (2003a) Hanley flax. Can J Plant Sci 83:85–87

Duguid SD, Kenaschuk EO, Rashid KY (2003b) Lightning flax. Can J Plant Sci 83:89–91

Duguid SD, Kenaschuk EO, Rashid KY (2003c) Macbeth flax. Can J Plant Sci 83:803–805

Duguid SD, Kenaschuk EO, Rashid KY (2004) Prairie blue flax. Can J Plant Sci 84:801–803

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407

Everaert I, Riek JD, Loose MD, Waes JV, Bockstaele EV (2001) Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum* L.) varieties by means of AFLP and morphological data. Plant Var Seeds 14:69–87

Feingold S, Lloyd J, Norero N, Bonierbale M, Lorenzen J (2005) Mapping and characterization of new EST-derived microsatellites for potato (*Solanum tuberosum* L.). Theor Appl Genet 111:456–466

Fraser LG, Harvey CF, Crowhurst RN, De Silva HN (2004) EST-derived microsatellites from *Actinidia* species and their potential for mapping. Theor Appl Genet 108:1010–1016

Fu Y, Peterson G, Diederichsen A, Richards KW (2002) RAPD analysis of genetic relationships of seven flax species in the genus *Linum* L. Genet Res Crop Evol 49:253–259

Fu Y, Rowland GG, Duguid SD, Richards KW (2003a) RAPD analysis of 54 North American flax cultivars. Crop Sci 43:1510–1515

Fu YB, Guerin S, Peterson GW, Diederichsen A, Rowland GG, Richards KW (2003b) RAPD analysis of genetic variability of regenerated seeds in the Canadian flax cultivar CDC Normandy. Seed Sci Technol 31:207–211

Fu YB, Peterson GW, Yu JK, Gao L, Jia J, Richards KW (2006) Impact of plant breeding on genetic diversity of the Canadian hard red spring wheat germplasm as revealed by EST-derived SSR markers. Theor Appl Genet 112:1239–1247

Gong L, Stift G, Kofler R, Pachner M, Lelley T (2008) Microsatellites for the genus *Cucurbita* and an SSR-based genetic linkage map of *Cucurbita pepo* L. Theor Appl Genet 117:37–48

Gorman MB, Cullis CA, Aldridge N (1993) Genetic and linkage analysis of isozyme polymorphisms in flax. J Hered 84:73–78

Han ZG, Guo WZ, Song XL, Zhang TZ (2004) Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Mol Genet Genomics 272:308–327

Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T (2006) Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theor Appl Genet 112:430–439

Hisano H, Sato S, Isobe S, Sasamoto S, Wada T, Matsuno A, Fujishiro T, Yamada M, Nakayama S, Nakamura Y, Watanabe S, Harada K, Tabata S (2007) Characterization of the soybean genome using EST-derived microsatellite markers. DNA Res 14:271–281

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Huang XQ, Cloutier S, Lycar L, Radovanovic N, Humphreys DG, Noll JS, Somers DJ, Brown PD (2006) Molecular detection of QTLs for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (*Triticum aestivum* L.). Theor Appl Genet 113:753–766

Kenaschuk EO, Rashid KY (1993) AC Linora flax. Can J Plant Sci 73:839–841

Kenaschuk EO, Rashid KY (1994) AC McDuff flax. Can J Plant Sci 74:815–816

Kenaschuk EO, Rashid KY (1998) AC Watson flax. Can J Plant Sci 78:465–466

Kenaschuk EO, Rashid KY (1999) Nouveau cultivar de lin AC Carnduff. Can J Plant Sci 79:373–374

Kenaschuk EO, Rashid KY, Gubbels GH (1996) AC Emerson flax. Can J Plant Sci 76:483–485

Krulickova K, Posvec Z, Griga M (2002) Identification of flax and linseed cultivars by isozyme markers. Biol Plant 45:327–336

La Rota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6:23

McBreen K, Lockhart PJ, McLenachan PA, Scheele S, Robertson AW (2003) The use of molecular techniques to resolve relationships among traditional weaving cultivars of *Phormium*. NZ J Bot 41:301–310

Mian MA, Saha MC, Hopkins AA, Wang ZY (2005) Use of tall fescue EST-SSR markers in phylogenetic analysis of cool-season forage grasses. Genome 48:637–647

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

Oh TJ, Gorman M, Cullis CA (2000) RFLP and RAPD mapping in flax (*Linum usitatissimum*). Theor Appl Genet 101:590–593

Park YH, Alabady MS, Ulloa M, Sickler B, Wilkins TA, Yu J, Stelly DM, Kohel RJ, el-Shihy OM, Cantrell RG (2005) Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Mol Genet Genom 274:428–441

Peng JH, Lapitan NL (2005) Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. Funct Integr Genomics 5:80–96

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238

Roose Amsaleg C, Cariou Pham E, Vautrin D, Tavernier R, Solignac M (2006) Polymorphic microsatellite loci in *Linum usitatissimum*. Mol Ecol Notes 6:796–799

Rowland GG, Hormis YA, Rashid KY (2002) CDC Bethune flax. Can J Plant Sci 82:101–102

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, pp 365–386

Rungis D, Berube Y, Zhang J, Ralph S, Ritland CE, Ellis BE, Douglas C, Bohlmann J, Ritland K (2004) Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theor Appl Genet 109:1283–1294

Saeidi G (2008) Genetic variation and heritability for germination, seed vigour and field emergence in brown and yellow-seeded genotypes of flax. Int J Plant Prod 2:15–22

Saha MC, Cooper JD, Mian MA, Chekhovskiy K, May GD (2006) Tall fescue genomic SSR markers: development and transferability across multiple grass species. Theor Appl Genet 113:1449–1458

Scott KD, Eggler 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

Spielmeyer W, Green AG, Bittisnich D, Mendham N, Lagudah ES (1998) Identification of quantitative trait loci contributing to Fusarium wilt resistance on an AFLP linkage map of flax (*Linum usitatissimum*). Theor Appl Genet 97:633–641

Stegnii VN, Chudinova YV, Salina EA (2000) RAPD analysis of flax (*Linum usitatissimum* L.) varieties and hybrids of various productivity. Genetika Moskva 36:1370–1373

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

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422

Tyson H, Fieldes MA, Cheung C, Starobin J (1985) Isozyme relative mobility changes relative to leaf position; apparently smooth trends and some implications. Biochem Genet 23:641–654

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

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

Varshney RK, Grosse I, Hahnel U, Siefken R, Prasad M, Stein N, Langridge P, Altschmied L, Graner A (2006) Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. Theor Appl Genet 113:239–250

Wiesner I, Wiesnerova D, Tejklova E (2001) Effect of anchor and core sequence in microsatellite primers on flax fingerprinting patterns. J Agric Sci 137:37–44

Xu Y, Ma RC, Xie H, Liu JT, Cao MQ (2004) Development of SSR markers for the phylogenetic analysis of almond trees from China and the Mediterranean region. Genome 47:1091–1104

Yu JK, Dake TM, Singh S, Benscher D, Li W, Gill B, Sorrells ME (2004) Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome 47:805–818

Zhang LY, Ravel C, Bernard M, Balfourier F, Leroy P, Feuillet C, Sourdille P (2006) Transferable bread wheat EST-SSRs can be useful for phylogenetic studies among the Triticeae species. Theor Appl Genet 113:407–418

Zohary D (1999) Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. Genetic Res Crop Evol 46:133–142