J. For. Res. (2016) 27(3):639–646
DOI 10.1007/s11676-015-0182-1

ORIGINAL PAPER

Analysis of EST-SSRs in silver birch (*Betula pendula* Roth.)

Ertugrul Filiz[1] · Ilhan Dogan[2,3] · Ibrahim Ilker Ozyigit[4]

Received: 14 July 2014 / Accepted: 11 March 2015 / Published online: 1 December 2015

Abstract Simple sequence repeats (SSRs) defined as sequence repeat units between 1 and 6 bp occur abundantly in both coding and non-coding regions in eukaryotic genomes and these repeats can affect gene expression. In this study, ESTs (expressed sequence tags) of *Betula pendula* (silver birch) were analyzed for *in silico* mining of EST-SSRs, protein annotation, open reading frames (ORFs), designing primers, and identifying codon repetitions. In *B. pendula*, the frequency of ESTs containing SSRs was 7.8 % with an average of 1SSR/4. 78 kb of EST sequences. A total of 188 SSRs was identified by using MISA software and di-nucleotide SSR motifs (65.9 %) were found to be the most abundant type of repeat motif followed by tri- (27.1 %), tetra- (4.8 %), and penta- (2.2 %) motifs. Based on ORF analysis, 175 of 178 sequences were predicted as ORFs and the most frequent SSRs were detected in 5′ UTR (58.43 %), followed by in ORF (31.46 %) and in 3′ UTR (8.43 %). 102 of 178 ESTs were annotated as ribosomal protein, transport protein, membrane protein, carrier protein, binding protein, and transferase protein. For a total of 102 SSRs (57.3 %) with significant matches, a set of 102 primers (100 %) with forward and reverse strands was designed by using Primer3 software. Serine (Ser, 19.6 %) was predominant in putative encoded amino acids and most of amino acids showed nonpolar (35.3 %) nature. Our data provide resources for *B. pendula* and can be useful for in silico comparative analyses of Betulaceae species, including SSR mining.

Keywords Silver birch (*Betula pendula*) · Betulaceae · EST-SSR · SSR mining · *In silico* analysis

Introduction

The genus *Betula* includes 30–60 species of trees and shrubs is widespread throughout the boreal and temperate climate zones of the Northern Hemisphere. The genus *Betula* belongs to the order Fagales and the Betulaceae family (Furlow 1990; De Jong 1993; Schenk et al. 2008). The basic chromosome number of *Betula* is n = 14, but the species form a series of polyploids, with chromosome numbers of 28, 56, 70, 84, 112, and 140 (Furlow 1990). The genus *Betula* is represented in Europe by two tree species: *B. pubescens* and *B. pendula*. *B. pendula* is distributed over most of Europe, south of Asia (Siberia, Iran, and Anatolia) and north of Africa (Martín et al. 2008).

Simple sequence repeats (SSRs) also known as microsatellites are 1–6 bp long tandemly repeating short units and are located in both coding and non-coding regions of all higher organism genomes (Tautz and Renz 1984; Gupta et al. 1996). Owing to their abundance and high mutation rate, they can be used as effective molecular

The online version is available at http://www.springerlink.com

Corresponding editor: Hu Yanbo

✉ Ertugrul Filiz
ertugrulfiliz@gmail.com

[1] Department of Crop and Animal Production, Cilimli Vocational School, Duzce University, Cilimli, 81750 Duzce, Turkey

[2] Department of Molecular Biology and Genetics, Izmir Institute of Technology, Urla, 35430 Izmir, Turkey

[3] Department of Biology, Faculty of Science, Rahatlayadururken Manas University, 72001 Bishkek, Kyrgystan

[4] Department of Biology, Faculty of Science and Arts, Marmara University, Goztepe, 34722 Istanbul, Turkey

markers, especially in genetic diversity and linkage mapping studies (Powell et al. 1996; Bérubé et al. 2007). ESTs being segments of expressed genes are fast accumulating in EST databases in large number of plant species due to intensive studies on genomics. Also, they provide some benefits due to having abundance, occurrence in gene-rich regions, and inherent advantages (Scott et al. 2000; Rungis et al. 2004). The EST databases can be used effectively for SSR mining (Varshney et al. 2002). EST–SSRs or genic SSRs as molecular markers can be obtained by database searches and other *in silico* approaches and can be used in transferability studies because they contain conserved genic regions (Varshney et al. 2002; Gupta et al. 2010a, b). Many EST-SSRs studies have been performed using various plant species, including the medicinal plant *Ocimum basilicum* (Gupta et al. 2010a, b), *Quercus robur* (Filiz et al. 2012), *Citrus sinensis* (Shanker et al. 2007), some cereal species (Varshney et al. 2002), loblolly pine and spruce (Bérubé et al. 2007), *Eucalyptus globules* (Acuña et al. 2011), *Ricinus communis* (Qiu et al. 2010). In this study, we performed in silico mining of EST-SSRs in silver birch (*B. pendula*) for analyses of SSR distributions and abundance, development of EST-SSR markers, prediction of open reading frames (ORFs) and annotation of SSR containing sequences.

Materials and methods

Retrieval and assembly of EST sequences

All EST sequences of *B. pendula* were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/nucest/). A total of 2549 ESTs were detected in the tissues (leaf, stem, root, etc.) of *B. pendula*. To remove redundant comparisons, CAP3 (sequence assembly program) was used with its default parameters (Huang and Madan 1999).

Identification of SSR motifs

Identification of SSR motifs was carried out by using MISA program (MIcroSAtellite) (http://pgrc.ipk-gatersleben.de/misa/) written in the Perl scripting language. The minimum length of SSR was accepted as 14 bp according to criteria used by Gupta et al. (2003). The SSRs were defined as ≥14 bp di-, ≥15 bp tri-, ≥16 tetra-, ≥20 penta-, and ≥24 hexa-nucleotide repeats.

SSR-EST similarity searches and functional annotation of significant matches

Pairwise comparison of SSR-EST sequences against the GenBank non-redundant protein database was done by using BLASTX program at NCBI database. The most significant matches (EXP $< 1e^{-6}$ and 70 % similarity) for each sequence were recorded.

Detection of SSR positions based on ORFs

ORFs were predicted for all SSRs containing sequences with ORF finder at NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) using standard genetic code. Uninterrupted by stop codon and maximum length were accepted as the primary encoding segment (ORF) for query sequences. All predictions were classified in three locations: within the ORF, in the 5′ untranslated region (UTR), or in the 3′ UTR (Shanker et al. 2007).

Primer designing

Primer sequence designing of SSR-EST sequences was performed with PRIMER3 software (http://frodo.wi.mit.edu/primer3/). The conditions for primer designing were adopted as default values. Also, the putative SSRs coding amino acids were determined and classified based on the physiochemical properties of amino acids.

Results

ESTs resource

We used a total of 2549 ESTs (987.395 bp) for SSR mining with a minimum length of 14 bp. The reduction in redundancy was used as a measure of degree of overlap among EST sequences (Gupta et al. 2010a, b). Based on assembled data (2278 ESTs with 899,028 bp), the percentage of ESTs forming contigs was 6.7 % (152) whereas 93.3 % of ESTs (2126) were unique and had no corresponding overlapping sequences. Thus, the reduction in redundancy was found to be 10.6 %.

Distribution of EST-SSRs

The screened genome data from *B. pendula* yielded a total of 188 for the presence of SSRs (Table 1), giving an average density of 1SSR/4.78 kb and only 7.8 % of assembled sequences contained SSRs.

The frequencies of SSR types with di-, tri-, tetra- and hexa-nucleotide repeat units are shown in Fig. 1. The most frequent repeat type was found to be as di-nucleotides (124, 65.9 %) followed by tri- (51, 27.1 %), tetra- (9, 4.8 %), and penta-nucleotides (4, 2.2 %). Interestingly, we identified no hexa-nucleotide repeat.

EST-SSRs were composed of seven different types of di-nucleotide: $(AG)_n$, $(AT)_n$, $(CT)_n$, $(GA)_n$, $(GT)_n$, $(TA)_n$, and $(TC)_n$, 16 different types of tri-nucleotide: $(AAC)_n$, $(AAG)_n$,

Table 1 Summary of EST-SSR mining of *B. pendula*

Parameters	Values
Total number of sequences examined	2278
Total size of examined sequences (bp)	899,028
Total number of identified SSRs	188
Number of SSR containing sequences	178
Number of sequences containing more than 1 SSR	10
Number of SSRs present in compound formation	9
Repeat type	
Dinucleotide	124
Trinucleotide	51
Tetranucleotide	9
Pentanucleotide	4

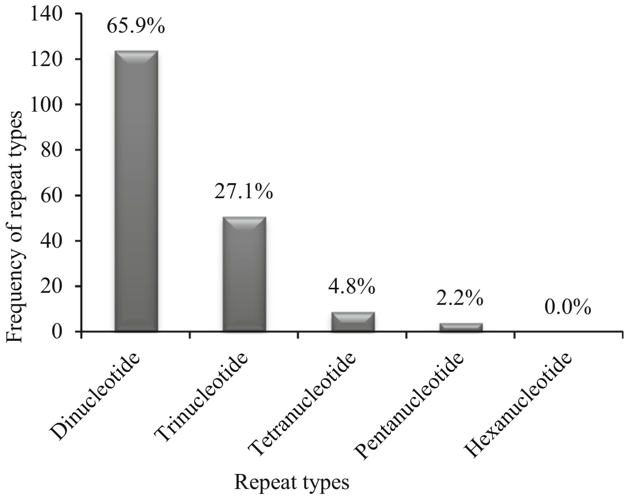

Fig. 1 Frequency distribution of different repeat types identified in ESTs of *B. pendula*

$(AGA)_n$, $(ATG)_n$, $(CAT)_n$, $(CTT)_n$, $(GAA)_n$, $(GAC)_n$, $(GAT)_n$, $(GGT)_n$, $(GTG)_n$, $(TCA)_n$, $(TCT)_n$, $(TGG)_n$, $(TTA)_n$, and $(TTC)_n$, four different types of tetra-nucleotide: $(CCGC)_n$, $(GAGC)_n$, $(GCGA)_n$, and $(GGAA)_n$, two different types of penta-nucleotide: $(ATCTG)_n$ and $(GATCC)_n$. Among di-nucleotide SSRs, GA motif types was the most abundant motif (20.7 %), followed by TC (14.4 %) and CT (12.7 %) motifs. Among tri-nucleotide SSRs, the most frequent motif was AAG (3.7 %), followed by GAT and TCA (3.2 %) and GTG (2.6 %). GAGC repeat type was the most frequent motif in tetra-nucleotides (3.2 %) whereas ATCTG and GATCC shared equal frequency at 1.1 % in penta-nucleotides.

Distribution of tri-nucleotide SSRs and putative encoded amino acids

Tri-nucleotide motifs code for corresponding amino acids and therefore play roles in determining biological activity of rotein molecules (Gupta et al. 2010a, b). Out of a total of 51 tri-nucleotides, 19.6 % of tri-nucleotides encoded serine, followed by aspartic acid and lysine shared in equal frequency at 13.7 % (Fig. 2). Putative encoded amino acids were grouped based on their polar and non polar nature, and according to the data, 35.3 % of amino acids were in non-polar nature, while 23.5 % were in polar nature.

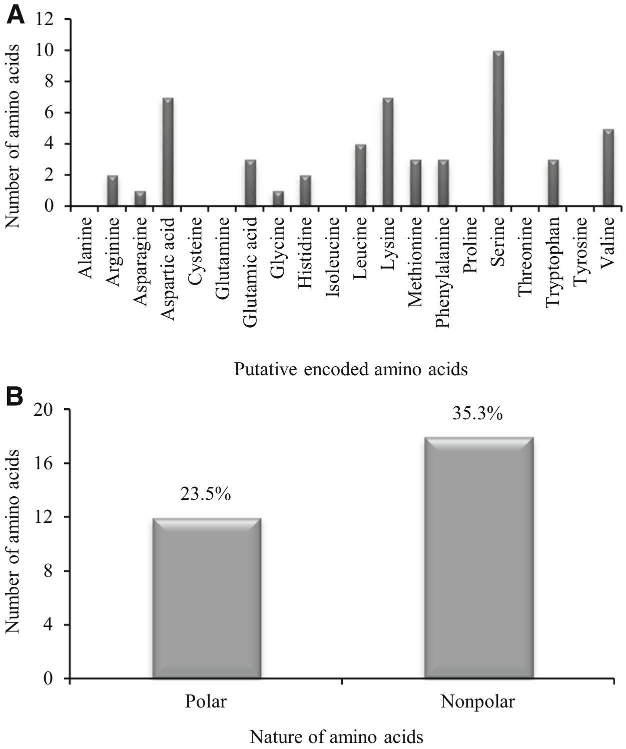

Fig. 2 Distribution of putative encoded amino acids (**a**), percentage frequency of polar and non-polar amino acids (**b**)

Analysis of BLASTX results

To determine the function of SSR containing sequences, 178 SSRs containing sequences were analyzed against the non-redundant (nr) protein database in NCBI (http://www.ncbi.nlm.nih.gov); thus, 102 of 178 EST-SSRs (57.3 %) were annotated. These proteins belong to non-specific lipid-transfer protein (8, 7.84 %), elongation factor 1-beta (6, 5.88 %), profilin (6, 5.88 %), 60S ribosomal protein (5, 4.90 %), 50S ribosomal protein (5, 4.90 %), small acidic protein (4, 3.92 %), catalase heme-binding enzyme (4, 3.92 %), alpha/beta-hydrolases superfamily protein (4, 3.92 %), small nuclear ribonucleoprotein (4, 3.92 %), ATP-dependent clp protease proteolytic subunit-related protein (3, 2.94 %), metallothionein-like protein (3, 2.94 %), acyl carrier protein (3, 2.94 %), NADH dehydrogenase [ubiquinone] 1 alpha sub-complex subunit 2 (3, 2.94 %), glycine-rich RNA-binding protein (2, 1.96 %), copper transport protein (2, 1.96 %), fructose-bisphosphate aldolase (2, 1.96 %), deSI-like protein (2, 1.96 %), heavy-metal-associated domain-containing family protein (2, 1.96 %), dof zinc finger protein (2, 1.96 %), ferredoxin-3 (2, 1.96 %), ubiquitin-related modifier (2, 1.96 %), zinc finger (C2H2 type) family protein (2, 1.96 %) splicing factor 3A subunit 2-like (1, 0.98 %), cold acclimation protein WCOR413 (1, 0.98 %), proteasome subunit alpha type (1, 0.98 %), nucleotide-diphospho-sugar transferase superfamily protein 14-3-3 protein 14-3-3 protein (1, 0.98 %), lipase class 3 family protein (1, 0.98 %), H/ACA ribonucleoprotein complex subunit 3-like protein (1, 0.98 %), snakin-2-like (1, 0.98 %), proteasome subunit alpha type (1, 0.98 %), auxin-repressed 12.5 kDa protein (1, 0.98 %), flavodoxin-like quinone reductase (1, 0.98 %), xanthoxin dehydrogenase-like (1, 0.98 %), CAX-interacting protein (1, 0.98 %), heat shock protein 90 (1, 0.98 %), FAS-associated factor (1, 0.98 %), 40S ribosomal protein (1, 0.98 %), cyclin-P3-1 (1, 0.98 %), mitochondrial outer membrane protein porin 2-like (1, 0.98 %), cytidine/deoxycytidylate deaminase family protein (1, 0.98 %), mitochondrial phosphate carrier protein (1, 0.98 %), aquaporin PIP2.2 (1, 0.98 %), and ADP-ribosylation factor-like protein (1, 0.98 %).

Primer designing for SSRs

Out of 102 EST-SSRs with significant matches, primers were designed for all EST-SSRs, including 98 singletons and 4 contigs (Table 2). These primers of 102 EST-SSRs include 59 di-, 30 tri-, 9 tetra-, and 4 penta-nucleotides.

Prediction of ORFs in SSR containing sequences

According to predicted ORFs analyses, a total of 175 EST-SSR containing sequences included ORFs whereas only three EST-SSRs had no ORF. The most frequent SSRs were predicted in 5′ UTR (104, 59.43 %), followed by in ORF (56, 32 %) and in 3′ UTR (15, 8.5 %) (Fig. 3).

Discussion

Microsatellites consist of repeated short nucleotide motifs (1–6 bp) (Tautz 1993). Owing to the mutation process known as DNA replication slippage, microsatellites gain and lose repeating units at high rates (Ellegren 2000). In this study, a total of 2549 ESTs from silver birch (*B. pendula*) were mined for simple sequence repeats. Also, protein annotations, primer designing, and prediction of ORFs were performed. The percentage of EST-SSR sequences reported here (7.8 %) is higher than the results of previous studies of other plant species. For example, those numbers were 3.60 % for barley (Kota et al. 2001), 2.80 % for sugarcane (Cordeiro et al. 2001), 2.70 % for cotton (Lü et al. 2010), and 5.70 % for *Lolium* (Asp et al. 2007), 6.62 % for *Quercus robur* (Filiz et al. 2012) and 1.1 % for both loblolly pine and spruce (Bérubé et al. 2007). A total of 188 EST-SSRs were detected with density of 1SSR/4.78 kb and our results showed lower values than reported in earlier studies: 1SSR/3.4 kb for rice (Varshney et al. 2002), 1SSR/1.67 kb for wheat (Morgante et al. 2002), 1 SSR/1.3 kb for *S. lycopersicum* (Gupta et al. 2010a), 1SSR/0.7 kb, 1SSR/1.67 kb, 1SSR/0.22 kb and 1SSR/3.5 kb for four different species including the sequences of major palms like coconut, arecanut, oil palm and date palm respectively (Palliyarakkal et al. 2011), 1SSR/1.77 kb for *R. communis* (Qiu et al. 2010). However, the SSRs density of *B. pendula* was higher than 1SSR/12.92 kb for *C. sinensis* (Shanker et al. 2007), 1SSR/6 kb for *Arabidopsis* (Cardle et al. 2000), 1SSR/12.92 kb for cotton (Lü et al. 2010), 1SSR/9.8 kb for *Q. robur* (Filiz et al. 2012), 1SSR/56.6 kb for loblolly pine and 1SSR/42.9 kb for spruce (Bérubé et al. 2007). Microsatellite genesis is due mainly to DNA slippage, unequal crossing over, gene conversion, and retro-transposition (Kalia et al. 2011). The frequencies of slippage and point mutations may affect distribution of SSRs in a genome. Also, the SSRs in genes contain higher mutation rates than non-repetitive regions (Li et al. 2004). The density and sequence numbers of EST-SSRs in *Betula* may be induced by mutation mechanisms and genome re-organization in evolutionary history.

In this study, the most abundant motif was found to be di-nucleotides (124, 65.9 %). This is in agreement with the results of earlier studies of *Arabidopsis* (Cardle et al. 2000), *C. sinensis* (Shanker et al. 2007), and loblolly pine and spruce (Bérubé et al. 2007). These similar results might be related to SSRs in non-coding regions. Three-nucleotide repeats are an abundant motif in *Betula* and similar results were found in earlier studies, e.g. *Lolium perenne* (Asp et al. 2007), castor bean (Qiu et al. 2010), medicinal plant *Ocimum basilicum* (Gupta et al. 2010a, b), and *Q. robur* (Filiz et al. 2012). Toth et al. (2000) and Morgante et al. (2002) reported a higher frequency of GA/CT repeat than AT repeat in *Arabidopsis* and cereals and our findings (GA was the most frequent motif in *Betula*) were similar to these study results. In three-nucleotides, AAG in *Arabidopsis*, CCG in cereals (Varshney et al. 2002), AAG/CTT in cotton (Lü et al. 2010), AAG/CTT in *Q. robur* (Filiz et al. 2012), AAG/CTT in castor bean (Qiu et al. 2010) were the most frequent motifs and the data supports our findings revealing AAG repeat (3.7 %) was the most abundant motif in *Betula*. The abundance of tri-nucleotide SSRs in coding regions may be related to the selective risk of non-trimeric SSR variants in coding regions, perhaps resulting from frame-shift mutations (Metzgar et al. 2000).

Table 2 Details of SSR containing ESTs with significant matches in *B. pendula*, including accession numbers, repeat motif, primer sequences, product size, and annealing temperature

No.	Accession number	Motif	Forward/reverse primer	Product size (bp)	Annealing temperature (°C)
1	CD278929	(AG)8	GTGCGACGAGACACAGAGAG/TAAGCCAGTTGCGATGTCAG	210	59.76
2	CD278925	(CT)12	CATTCCTCGAATGGATTGCT/AGAAAAGGCGCAACACAACT	195	59.92
3	CD278888	(CT)10	GCAGATTGAACACGCTTTGA/TTTTTCGAACCAAAACTCGAA	209	60.00
4	CD278876	(CT)7	TTCAATCTTTAGGCGCGTTT/TAAAGGGCATTCCACAGGTC	208	59.85
5	CD278819	(TCA)5	CGTCATCTCCGTGGACAATA/CAACCAGCCATACACACCAT	202	59.29
6	CD278757	(AAG)6	CGACTCTGTGGAGTCATGGA/CGCCTTTATCTCCAGCTTTG	211	59.82
7	CD278750	(TCA)5	GGACTTCTTCGGAGACATGG/AATGTCGGAGTCGTCGAAGT	209	59.73
8	CD278749	(ATCTG)4	GGCGATATCCTTGAATCGAA/GTCTCGCGGTCGTTGATAAT	194	60.10
9	CD278733	(AT)9	AAAACCGTGTCTCTCGCAGT/CCGGAAAATTTGGTCCACTA	205	59.79
10	CD278714	(AG)10	TCTACGCAGACGCAGAGAGA/AGGAGGTTCCTTTCCTCCAC	199	59.53
11	CD278710	(TCA)5	CGTCATCTCCGTGGACAATA/AACCAGCCAAACACACCATT	201	59.52
12	CD278665	(GCGA)4	GAGCGAGCGAGAGACAGACT/CCCCTTTCTTCAGATCAACG	206	59.66
13	CD278629	(CT)7	GCTCAAGTGGGTCCGTTTTA/ACCGCAGAAGCATACTCCAC	204	60.29
14	CD278585	(TC)8	TCAATGAGAATGGCGACAAA/GAATGCCGCAAGAGTTCAAT	191	60.22
15	CD278562	(GAT5)	TTCAGTTGATGCTGCTCCTG/TCTCATCATCCCAAGGCTTC	203	60.16
16	CD278540	(CCGC)4	TGTGGAGGAGCATGTAGTGC/GGGCGGACATATGTTTCAAG	195	59.86
17	CD278539	(CT)9	TGGGTCCAGACTTTCGAGTT/CATTGTCATCAAACCCAGCA	192	59.70
18	CD278525	(TA)7	GTAGGCCGCCCAAAGACTAT/CGTTTCGTCCCAAATCTTGT	170	59.97
19	CD278521	(GAT)5	AGACTCCTTTCCGAATGCAA/CGTCCTCTGTCTCATCACCA	200	59.82
20	CD278493	(GAGC)5	GAGCGAGCGAGAGACAGACT/CCCCTTTCTTCAGATCAACG	206	59.66
21	CD278458	(AT)8	CTCGACATGGGCTACACAGA/CAAGGCCTTCTCTGCTTCAA	194	59.86
22	CD278435	(AAG)5	GAAGATGTCGTGGCAAACCT/ACCAAGGTACAGGCCAGTTG	208	60.03
23	CD278413	(TC)7	GGCTCTTCCTGCTGATTCTG/CACTCCCCTTCTTCTCAACG	200	59.84
24	CD278341	(GAGC)5	GAGCGAGCGAGAGACAGACT/CCCCTTTCTTCAGATCAACG	206	59.66
25	CD278329	(AG)7	TGGAGGGGTCATCTATCTCG/CCCCTCTCAACACCCATATC	206	59.21
26	CD278320	(CT)7	CAAACCAGAGCCTCTCATCC/GCGTCCCTATAGGCATCATT	202	59.03
27	CD278298	(TC)13	TGATAGGGAGCGGATACCAG/ACGAGGATCCCTCAAGGTTT	199	59.93
28	CD278287	(ATG)5	AAGATCGCCTTCGATGACAC/CCAATCGCCTACTTGACCAC	199	60.52
29	CD278258	(AG)10	AGAGCCGCTTCACAGAGAGA/GGTTTCCTCGTCGGTTATGA	199	59.93
30	CD278220	(AG)12	GAAAGGGGCTTCATCAGTTG/TTTGAATTGAGCAGCCATCA	199	59.67
31	CD278175	(AAG)5	GAAGATGTCGTGGCAAACCT/ACCAAGGTACAGGCCAGTTG	208	60.12
32	CD278136	(AG)7	TGGAGGGGTCATCTATCTCG/TGGAGGGGTCATCTATCTCG	206	59.21
33	CD278102	(TGG)5	ATGAAGGTGATCGCTGCATA/CCAGAAGGCACAGATGCTAA	203	59.26
34	CD278084	(TC)11	GCGACAGGAAATTCAACCAC/ACGTTCTGCTCCTTCAATCG	211	60.40
35	CD278049	(TGG)6	TGCTCTCGTTTCCAACCTCT/CCCTTGACTTCACACAAGAGC	195	59.90
36	CD277996	(CT)10	AAACATGTCGGCTTTCAAGG/AAACATGTCGGCTTTCAAGG	200	60.66
37	CD277982	(AG)8	AGAGCAAGCCGAGAGATACG/CTCCTCGCATTCTGTTCGTT	201	59.74
38	CD277916	(GAGC)5	GAGCGAGCGAGAGACAGACT/CCCCTTTCTTCAGATCAACG	206	59.66
39	CD277897	(AG)8	AAGTGGCATTGGTCACAGGT/TGCACGGCATACATCATCTT	197	60.43
40	CD277853	(GAT)5	CCTCGACGAGTTCCTCTCTG/TGAAGCACCTTTGCCATACA	205	60.26
41	CD277850	(CT)14	TGATAGGGAGCGGATACCAG/ACGAGGATCCCTCAAGGTTT	199	59.93
42	CD277824	(TC)8	TTTTGGGCTCCATCTCATTC/TTGAGTCCCGTCCATTCTTT	202	59.53
43	CD277797	(AG)8	GAAAGGGAAGCGAAGAGGAC/TCCTTATTGGGTGCATAGGG	190	59.78
44	CD277764	(AAG)5	GAAGATGTCGTGGCAAACCT/ACCAAGGTACAGGCCAGTTG	208	60.03
45	CD277747	(GA)8	GACTACCTCCGCTTCGTCAC/GAGCTACCTCGTCGTCGTTG	202	59.87
46	CD277714	(GAGC)5	GAGCGAGCGAGAGACAGACT/CCGATTTTCTTAGGAGCGATG	214	61.05

Table 2 continued

No.	Accession number	Motif	Forward/reverse primer	Product size (bp)	Annealing temperature (°C)
47	CD277683	(TCA)5	CGTCATCTCCGTGGACAATA/AACCAGCCAAACACACCATT	200	59.52
48	CD277622	(GAA)5	GAGCAGTGCCGGTTTATCTC/TGAATGTGAACCCAGGACAA	189	59.94
49	CD277609	(TC)7	GGCTCTTCCTGCTGATTCTG/CACTCCCCTTCTTCTCAACG	200	59.84
50	CD277589	(TC)13	GCAAGCGTCTTTCAAGCATT/CCTTGATACCAGGGAGAACG	201	59.54
51	CD277535	(CT)8	TTGGTGGTTGTTCTTGTCCA/CTCACAGACCCACTTGCAGA	204	59.98
52	CD277532	(AAG)5	GAAGATGTCGTGGCAAACCT/ACCAAGGTACAGGCCAGTTG	208	60.03
53	CD277530	(GAGC)4	GAGCGAGCGAGAGACAGACT/CCCCTTTCTTCAGATCAACG	206	59.66
54	CD277502	(TC)7	ACGCTTTCGTGTTTCTTGCT/TTCTTCTTCCACGCTGATCC	199	60.34
55	CD277473	(CT)9	CCAACAGGCTTTCATTTGCT/ACAAGAGCTCGGTTCTGGTC	200	59.45
56	CD277411	(CTT)5	GAGATGGCGGAGACTGAGAC/TGGATGAAAAGCACAGGTTG	200	59.69
57	CD277408	(AT)9	TGTAGCAGAGATGGCCTTGA/GGAATCCATGGCAAACCTTA	199	59.76
58	CD277395	(TC)8	TTTTGGGCTCCATCTCATTC/TTGAGTCCCGTCCATTCTTT	202	59.53
59	CD277391	(TGG)5	TGCTCTCGTTTCCAACCTCT/CCCTTGACTTCACACAAGAGC	195	59.90
60	CD277387	(CT)8	TGGGTGCTCTGGACTTTCTC/TACTGTTACCCGGCTCTGCT	194	59.90
61	CD277383	(AG)7	AGCTTCGTTCCAAAACCTCA/CCCATTTGGAGATGGAGAAA	208	59.86
62	CD277382	(TC)11	CGCAGAGTCTTCGACATGAG/TCACCACCATGCCAATATCA	199	60.76
63	CD277308	(GGT)5	GACATGCAATCCCTTGGAGT/ACAACTTGGGGTGGAAACAC	202	59.72
64	CD277302	(CT)11	AGAACTGTTGGGAGGTGCAG/CGTGGCTGAGTGAGGTTGTA	204	59.90
65	CD277285	(TC)11	CATGGTGGTGATCAGAGGAA/GGGCTAAAAATGGTCCACCT	197	60.19
66	CD277250	(CT)12	CGAATTGAAGGAGCAGAAGG/TGCTCACAGCAAAGCAGAGT	189	59.93
67	CD277249	(GATCC)4	CCTCCACCCGTTCAAGTGTA/TCACTTTGTGCACTGCCATT	210	60.31
68	CD277243	(TC)9	CATTCCAGTCCATTCCGTTC/CCAATTTGTCAGCCGTATCA	203	59.54
69	CD277238	(TCA)5	CGTCATCTCCGTGGACAATA/AACCAGCCAAACACACCATT	201	59.52
70	CD277237	(ATG)6	AAGATCGCCTTCGATGACAC/CCAATCGCCTACTTGACCAC	199	60.52
71	CD277232	(CT)14	GACTACCAACTCCGGTGCTC/TCATGGGTGACCTCAAAGAA	189	59.06
72	CD277148	(AG)10	GGAGTACAGGCAGAGGGTTG/CAGTGACTGATCCCCAGCTT	199	59.72
73	CD277140	(CT)8	TGGGTGCTCTGGACTTTCTC/TACTGTTACCCGGCTCTGCT	194	59.90
74	CD277125	(CT)11	AGAACTGTTGGGAGGTGCAG/CGTGGCTGAGTGAGGTTGTA	204	59.90
75	CD277116	(CT)7	AATTCGGTGGGGGACTAGAG/CATTCACAAGGACCAGAACG	195	59.13
76	CD277113	(TC)9	GCCGCTTTGAGACTCTGATT/GGGACATAGGTTGCATGCTT	202	59.96
77	CD277098	(GAA)5	GGCGTTTAATCTGGGTGAGA/TTCCTGATGTCGAATGCTCA	213	60.35
78	CD277085	(GA)9	CTGGCAAAAGTGTGCAGAAA/CTGAGCATCAAGTGCCAAAG	191	59.59
79	CD277013	(AG)11	TGGTTGAAGCGATGAAGACA/CAACCCTCTGCCTGTACTCC	201	59.72
80	CD276959	(TC)10	AATTCGGTGGGGGACTAGAG/CATTCACAAGGACCAGAACG	195	59.13
81	CD276953	(CT)8	TCCATTTCCTCAACCCTCTC/TGAGAGCCCTCCTTTCTTCA	199	59.06
82	CD276933	(TC)7	TGCTCCGGTTTACAACAATG/CGAAGCGATAGTGACGACAG	192	59.62
83	CD276907	(CT)8	TTGGTGGTTGTTCTTGTCCA/CTCACAGACCCACTTGCAGA	204	59.98
84	CD276864	(GAT)5	TTCAGTTGATGCTGCTCCTG/TCTCATCATCCCAAGGCTTC	203	60.16
85	CD276856	(CT)9	AGCGCTCCCTCTCTCTCTCT/TTAGCTCCCACCACGTTAGC	205	60.27
86	CD276840	(GATCC)4	CAGCTCCTTTGGACTCTTCG/TAAGGCACACGATCTGCTTG	183	60.01
87	CD276777	(TCT)6	CGGCTTTCTCTCCCTCTCTT/ATGACCACAGCGAACACTTG	204	59.75
88	CD276759	(TC)10	CGCAGAGTCTTCGACATGAG/TCACCACCATGCCAATATCA	199	59.73
89	CD276730	(CT)7	TTGGTGGTTGTTCTTGTCCA/CTCACAGACCCACTTGCAGA	204	59.98
90	CD276682	(ATCTG)4	CCCTATGGCGATATCCTTGA/GTCTCGCGGTCGTTGATAAT	200	59.88
91	CD276680	(GAT)5	CCTCGACGAGTTCCTCTCTG/TGAAGCACCTTTGCCATACA	205	60.26
92	CD276664	(TCT)6	CGCCAAATCTTTACCCAGAA/CATCTCGATCCTCTCCTTCG	208	59.90
93	CD276628	(CT)8	GGTGACTTGGTGGGTGTTCT/CCCACTTGCAGACAGCAATA	202	59.86

Table 2 continued

No.	Accession number	Motif	Forward/reverse primer	Product size (bp)	Annealing temperature (°C)
94	CD276623	(TC)7	AGTGCTATGGCGAAGGACAT/TCGGACTGGCTCTTGTATCC	206	59.72
95	CD276610	(TCT)6	TCGCTGTGGTCATGATAGGA/ATCTTTGCCCTGCTCTTCAA	206	59.96
96	CD276602	(GGAA)4	GGTTTGTGCCAACGAAACTT/GGCAAGACCTTCCTTCCTTC	197	60.19
97	CD276578	(AT)8	CTCGACATGGGCTACACAGA/GTCAAGGCCATCTCTGCTTC	196	59.96
98	CD276494	(TCT)6	CGGCTTTCTCTCCCTCTCTT/ATGACCACAGCGAACACTTG	204	59.75
99	Contig 8	(AAG)5	GAAGATGTCGTGGCAAACCT/ACCAAGGTACAGGCCAGTTG	208	60.03
100	Contig29	(GAGC)4	TGGAGCAGATGAAGCAACAC/TGGAGCAGATGAAGCAACAC	203	59.97
101	Contig94	(GAT)5	CCTCGACGAGTTCCTCTCTG/TGAAGCACCTTTGCCATACA	205	60.26
102	Contig102	(AAG)5	GAAGATGTCGTGGCAAACCT/ACCAAGGTACAGGCCAGTTG	208	60.12

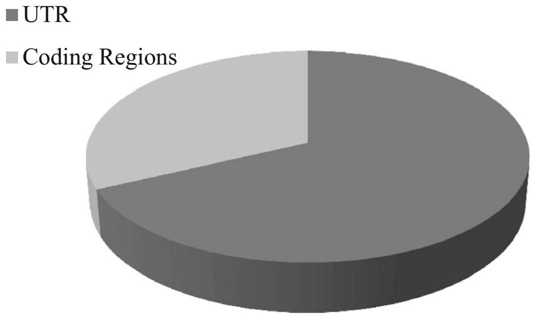

Fig. 3 Distribution of EST-SSRs in coding regions and UTRs

It was observed that serine (19.6 %) was found to be the most frequent amino acid encoded by trinucleotide SSRs. The most common amino acid was found to be serine (Ser) in *Quercus* and *Arabidopsis* (Lawson and Zhang 2006; Filiz et al. 2012) and these data are in agreement with our results. However, lysine (Lys) in *Arabidopsis* (Morgante et al. 2002) and arginine (Arg) in sugarcane (Cordeiro et al. 2001) were found to be the most encoded amino acids and these results contradict our findings. SSRs in promoter and intronic regions may affect gene activity and gene transcription. Also, SSR repeat numbers may regulate gene expression and expression level (Li et al. 2002). Lately, new evidence shows that large numbers of SSRs are located in coding regions of genomes (Morgante et al. 2002), suggesting that the difference in the distribution of amino acids could be a result of different gene expression profiles that affect SSRs in coding regions. In *Arabidopsis*, high densities of SSRs were found to be in UTRs (Lawson and Zhang 2006). Similar results were reported for the medicinal plant *Ocimum basilicum* (Gupta et al. 2010a, b) and for *C. sinensis* (Shanker et al. 2007). These data corroborated our results that nearly 68 % of SSRs were located in UTR regions in *Betula*. It can be proposed that SSRs in UTR regions play important roles in gene regulation. 102 of 178 EST-SSRs (57.3 %) were annotated and categorized into different classes, including ribosomal protein, transport protein, membrane protein, carrier protein, binding protein, transferase protein, and others. Also, 102 primers were designed for these putative proteins. In conclusion, ESTs of *B. pendula* were mined for EST-SSRs and the current study demonstrated that understanding of SSR distribution could support future studies with in silico comparative analysis of Betulaceae species, especially *Betula* taxa. Also, EST-SSR resources developed from this work can be utilized in studies of genetic diversity, linkage mapping, and comparative analyses in Betulaceae species.

References

Acuña CV, Fernandez P, Villalba PV, García MN, Hopp HE, Poltri SNM (2011) Discovery, validation, and in silico functional characterization of EST-SSR markers in *Eucalyptus globules*. Tree Genet Genomes. doi:10.1007/s11295-011-0440-0

Asp T, Frei UK, Didion T, Nielsen KK, Lübberstedt T (2007) Frequency, type, and distribution of EST-SSRs from three genotypes of *Lolium perenne*, and their conservation across orthologous sequences of *Festuca arundinacea*, *Brachypodium distachyon*, and *Oryza sativa*. BMC Plant Biol 7:36. doi:10.1186/1471-2229-7-36

Bérubé Y, Zhuang J, Rungis D, Ralph S, Bohlmann J, Ritland K (2007) Characterization of EST-SSRs in loblolly pine and spruce. Tree Genet Genomes 3:251–259

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

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

De Jong PC (1993) An introduction to *Betula*: its morphology, evolution, classification and distribution with a survey of recent work. The IDS Betula Symposium, International Dendrology Society, Susses

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 16:551–558

Filiz E, Koc I, Sakinoglu FC (2012) *In silico* EST-SSRs analysis in unigene of *Quercus robur* L. Res Plant Biol 2:1–9

Furlow JJ (1990) The genera of Betulaceae in the southeastern United States. J Arnold Arbor 71:1–67

Gupta PK, Balyan HS, Sharma PC, Ramesh B (1996) Microsatellites in plants: a new class of molecular markers. Curr Sci 70:45–54

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) EST-SSRs for transferability, polymorphism and genetic diversity in bread wheat. Mol Genet Genom 270:315–323

Gupta S, Shukla R, Roy S, Sen N, Sharma A (2010a) *In silico* SSR and FDM analysis through EST sequences in *Ocimum basilicum*. POJ 3:121–128

Gupta S, Tripathi KP, Roy S, Sharma A (2010b) Analysis of unigene derived microsatellite markers in family Solanaceae. Bioinformation 5:113–121

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

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. Euphytica 177:309–334

Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas 135:145–151

Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol 7:R14

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

Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21: 991–1007

Lü YD, Cai CP, Wang L, Lin SY, Zhao L, Tian LL, Lü JH, Zhang TZ, Guo WZ (2010) Mining, characterization, and exploitation of EST-derived microsatellites in *Gossypium barbadense*. Chin Sci Bull 55:1889–1893

Martín C, Parra T, Clemente-Muñoz M, Hernandez-Bermejo E (2008) Genetic diversity and structure of the endangered *Betula pendula* subsp. *fontqueri* populations in the south of Spain. Silva Fenn 42:487–498

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72–80

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

Palliyarakkal MK, Ramaswamy M, Vadivel A (2011) Microsatellites in palm (Arecaceae) sequences. Bioinformation 7:347–351

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

Qiu L, Yang C, Tian B, Yang JB, Liu A (2010) Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). BMC Plant Biol 10:278

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

Schenk MF, Thienpont CN, Koopman WJM, Gilissen LJWJ, Smulders MJM (2008) Phylogenetic relationships in *Betula* (Betulaceae) based on AFLP markers. Tree Genet Genomes 4:911–924

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

Shanker A, Bhargava A, Bajpai R, Singh S, Srivastava S, Sharma V (2007) Bioinformatically mined simple sequence repeats in UniGene of *Citrus sinensis*. Sci Hortic 113:353–361

Tautz D (1993) Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In: Pena SDJ, Chakraborty R, Epplen JT, Jeffreys AJ (eds) DNA fingerprinting: state of science. Birkhäuser Verlag, Basel, pp 21–28

Tautz D, Renz M (1984) Simple sequence repeats are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12:4127–4138

Toth G, Gáspári Z Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967–981

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