# Development of Juglans Regia SSR Markers by Data Mining of the EST Database

Rui Zhang · AnDan Zhu · XinJian Wang · Jun Yu ·
HongRong Zhang · JiangSheng Gao ·
YunJiang Cheng · XiuXin Deng

Published online: 13 March 2010
© Springer-Verlag 2010

**Abstract** Walnut (*Juglans regia*), an economically important woody plant, is widely cultivated in temperate regions for its timber and nutritional fruits. Despite abundant studies in germplasm, systemic molecular evaluations of walnut are sparsely reported mainly due to the limited molecular markers available. Expressed sequence tags (EST) provide a valuable resource for developing simple sequence repeat (SSR) markers. In this study, a total of 5,025 walnut ESTs (covering 16.41 Mb) were retrieved from the National Center for Biotechnology Information database. The SSR motifs were then analyzed by the SSRHunter software. In total, 398 SSRs were obtained with an average frequency of 1/4.08 kb. Dinucleotide (di-) repeat motifs accounted for 69.85% of all SSRs, followed by trinucleotide (tri-) with a frequency of 27.64%, while low frequency (2.51%) of tetranucleotide (tetra-) to hexanucleotide (hexa-) was observed. Meanwhile, GCA and TC motifs were prevalent among di- and tri- loci, respectively. Subsequently, a total of 123 primer pairs were designed from the non-redundant SSR-containing unigenes with the selection threshold of SSR length set to 10 bp or more. To examine the efficiency of candidate markers, seven DNA pools were collected from geographically different accessions. Results demonstrated that 41 SSR primer sets could generate high polymorphic amplification products (33.3%), and these polymorphic loci were mainly located in the 3′-untranslated region. Annotation analysis revealed that only two of these 41 loci were located inside open reading frames of characterized proteins ($E \leq 1E-30$).

**Keywords** *Juglans regia* · EST data mining · SSR primer · EST data analysis · Primer design

## Introduction

Simple sequence repeat (SSR), an array of short motifs of 1–6 bp in length, are hypervariable and widely spread in both coding and non-coding regions of plant and animal genomes (Kota et al. 2001). The numbers of core repeats are variable, presumably due to strand slippage during DNA replication or unequal exchange in meiosis. Because of the high mutability, SSR located regions are thought to play significant roles in genome evolution by creating and maintaining quantitative genetic variations (Fraser et al. 2004). The reproducibility, multiallelicism, codominance, relatively abundance, and good genome coverage of SSR markers have made them one of the most useful tools for integration of the genetic, physiological, and sequence-based physical maps in plant species (Aggarwal et al. 2007; Kota et al. 2001; Powell et al. 1996). However, the traditional genomic library-dependent approach for SSR markers development is time consuming and expensive. In the past decade, progresses on expressed sequence tag (EST)

**Electronic supplementary material** The online version of this article (doi:10.1007/s11105-010-0192-2) contains supplementary material, which is available to authorized users.

R. Zhang · X. Wang · J. Yu · J. Gao
Biological Technology Research & Development Center,
Tarim University,
Alaer, Xinjiang Uygur Autonomous Region 843300,
People's Republic of China

R. Zhang · A. Zhu · Y. Cheng (✉) · X. Deng
National Key Laboratory of Crop Genetic Improvement,
Huazhong Agricultural University,
Wuhan, Hubei 430070, People's Republic of China
e-mail: yjcheng@mail.hzau.edu.cn

R. Zhang · A. Zhu · H. Zhang · Y. Cheng · X. Deng
College of Horticulture and Forestry Science, Huazhong
Agricultural University,
Wuhan, Hubei 430070, People's Republic of China

sequencing projects have led to successive explosion of the EST database and its application. The online database offers an opportunity to identify simple sequence repeats in ESTs through data mining and provides a feasible and cost-efficient way for the development of SSR markers in plants (Saha et al. 2004). EST-SSR markers have been reported in many plant species, such as rye (Hackauf and Wehling, 2002), barley (Thiel et al. 2003), *Medicago truncatula* (Eujayl et al. 2004), tea plant (Jin et al. 2006), and wheat (Gao et al. 2003).

Walnut (*Juglans regia*) is an economically important plant for both timber use and its nutritional fruits. Although it has a long cultivation history and luxuriant wild species, genetic studies on walnut is chronically lagging behind compared with other fruit crops. To date, publicly available molecular markers for walnut such as SSR is limited, which significantly hampered the construction of high-density genetic maps in walnut (Dangl et al. 2005). Therefore, the present study was undertaken with the objective to develop and characterize a collection of EST-derived SSR markers for walnut germplasm evaluation.

## Materials and Methods

### Walnut EST Data Retrieval and SSR Detection

A total of 5,025 walnut EST sequences were retrieved from the NCBI website (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) on the third of April 2008. After removing vectors and contaminated sequences, raw sequences were assembled to eliminate redundancy using SeqMan module of the DNAStar software (Steve Shear Down 1998–2001 version reserved by DNASTAR Inc., Madison, WI) and submitted to the SSRHunter 1.3 software (Li 2003) for mining SSR motifs that were defined as at least five repeat units per motif.

### EST-SSR Primer Development

Primer pairs were designed from non-redundant unigenes with SSR motif using the Primer Premier 5.0 software (http://www.PremierBiosoft.com). The major parameters for primer design were GC content 40–60%, optimum annealing temperature 55°C, and PCR product size 100–400 bp in length. Primer pairs were synthesized by Shanghai Sangon Biological Engineering Technology (Shanghai, China).

To identify putative functions of the EST-SSR-containing genes, corresponding SSR unigenes were compared with the UniProtKB database (http://www.uniprot.org/) using the BLASTX program, assuming that an *E* value of ≤1E−5 was a significant homology criterion.

The POPGENE 1.31 software (Yeh and Boyle 1997) recognized marker types and estimated the codominant nature of SSR markers. The SSR marker diversity was estimated using the following parameters: observed number of alleles (na), effective number of alleles (ne), average expected heterozygosity (He), observed heterozygosity (Ho; Nei 1973), Shannon's Information index (*I*), and fixation index (Fst).

### Plant Materials and DNA Extraction

To experimentally verify the amplification specificity and polymorphism of the synthesized SSR primers, young leaves of 98 morphologically different plant accessions distributed in Xinjiang (Yecheng, Hutian, and Wensu counties), Tibet, Shandong, and Shannxi provinces were collected for DNA extraction. Wingnut (*Pterocarya hupehensis* Skan) was clustered as an outgroup because of its close genetic relationship with *Juglan J. regia*. DNA extraction was performed as described previously (Cheng et al. 2003). Subsequently, six DNA pools were constructed based on geographically different walnut accessions. An independent DNA pool was constructed with wingnut samples to serve as the control for revealing the maximal diminished polymorphic SSR alleles via PCR amplification. Therefore, a total of seven DNA pools was constructed and applied to subsequent experiments.

### PCR Amplification and Denatured Polyacrylamide Gel Analyses

PCR was carried out in a 20-μl reaction mixture containing 25 ng DNA as template, 0.1 μM of forward and reverse primers, 2.5 mM $MgCl_2$, 2.5 mM dNTPs, 1×*Taq* buffer (100 mM Tris–HCl pH8.0, 500 mM KCl, 0.8% Nonidet P40), and 1 U *Taq* DNA polymerase (Fermentas Inc., MD, USA). Amplification was performed in a PTC-200 DNA Engine Thermal Cycler (Bio-Rad Laboratories Inc., Hercules, California, USA) in 0.2-ml tubes.

Cycles were programmed as follows: initial denaturation cycle at 94°C for 3 min, 10 cycles of 30 s denaturing at 94°C, 30 s annealing at 66°C, 45 s elongation at 72°C; and 30 cycles of denaturing at 94°C, 30 s annealing at 55°C, 1 min elongation at 72°C, at last one final cycle of 5 min at 72°C, stored at 4°C. The PCR products were separated on 6% polyacrylamide denaturing gel and visualized by silver staining.

## Results

### Walnut EST-SSR Distribution, Frequency, and Other Features

A total of 5,025 walnut ESTs, which were sequenced from cDNA libraries of seed coats, mid-season walnut embryos and pellicles, were retrieved from NCBI database on third of April 2008 for subsequent data mining. After removing the redundant and junk sequences, 2,414 unigenes including 816 contigs and 1,598 singletons were finally obtained. In total, 398 potential SSR loci were subsequently identified, which represented 7.92% of all ESTs and 16.49% of the unigenes. The density of SSR distribution among walnut transcripts was approximately 4.18 kb per locus. The average number of repeats across all motifs was 6.77. The length of SSR ranged from 10 to 58 bp with an average of 15 bp. Frequent short SSRs in EST sequences were easily detected, while long SSRs (≥20 bp) only represented 22.9% of the motifs. The longest di-repeat motif was TC with a length of 58 bp (located in contig-796), and tri-repeat motif was GAT with a length of 39 bp (located in contig-81). Among the SSR loci, 54 compound types were detected, 44 of which were found with two immediately adjacent SSRs, nine with three SSRs, and only one unigene with four SSR loci. In addition, adjacent repeats with a space less than 10 bp were detected in 13 unigenes.

In the present investigation, the analyzed motifs ranging from di- to hexa- were divided into 65 repeat motif types. Eleven di- and 44 tri-motifs were obtained, which accounted for 84.62% (Fig. 1) of all types. This result demonstrated that nonrandom distribution of SSR unit sizes existed in walnut EST databases. SSR motifs were classified according to the repeat unit size. As a result, 278 di-SSR loci were preponderantly detected with a ratio and density of 69.85% and one SSR per 5.9 kb respectively; followed by 110 tri-loci with a ratio and density of 27.64% and one SSR per 14.92 kb. Low frequency was observed for tetra- to hexa-SSR loci. Only ten such loci were detected from the retrieved EST sequences, with the density around one SSR per 164.08 kb (2.51%; Table 1). The composition and distribution of the di- and tri-repeats were then analyzed. In dinucleotide repeats, results showed that TC repeats were predominant (114 loci), followed by GC and CA repeats. GCA motif was common in the tri-repeats.

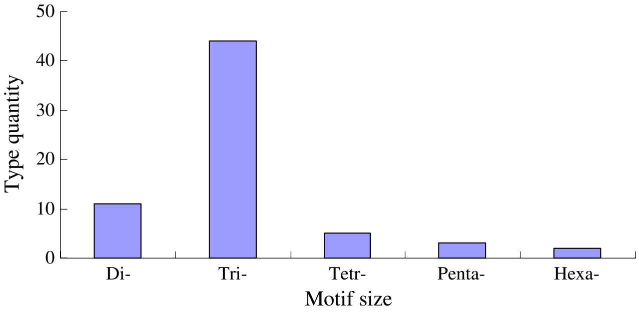

**Fig. 1** Frequencies of EST-derived SSRs in walnut genome based on repeat units: dinucleotide (*Di-*), trinucleotide (*Tri-*), tetranucleotide (*Tetra-*), pentanucleotide (*Penta-*), and hexanucleotide (*Hexa-*)

**Table 1** Characterization of the retrieved microsatellites

| Repeat type | Counts | Total loci | Proportion to all SSR loci% | Proportion to all EST % | Average SSR distribution (1/kb) |
|---|---|---|---|---|---|
| Di- | 11 | 278 | 69.85% | 5.53% | 5.90 |
| Tri- | 44 | 110 | 27.64% | 2.19% | 14.92 |
| Tetra-hexa- | 10 | 10 | 2.51% | 0.20% | 164.08 |
| Total | 65 | 398 | | | |

## Polymorphism Verification

After removing the unsuitable loci with short flanking sequences, a total of 123 primer pairs were designed from the 398 microsatellites. These primers were then synthesized for detection of seven walnut DNA pools. Ultimately, 41 out of the 123 primer pairs generated high-quality polymorphisms by PCR amplification using walnut genomic DNA as templates (see Table 1 of the Electronic Supplementary Materials). Twenty-six primer pairs could poorly amplify the targeted region, and smeared PCR products were seen on polyacrylamide gels. Forty primer pairs completely failed in amplification, and the remaining 16 primer pairs only amplified one single band from the tested samples (Fig. 2).

## Sequence Identification

Among the 41 polymorphic SSR loci, we found only two SSRs located inside ORFs, two located in 5′ UTRs (untranslated region), eight located in 3′ UTRs, and two at unknown positions when the $E$ value was set to 1E−30 for BLAST protein identification. Twenty-six SSRs remained either unassigned or not meeting the threshold of $E \leq 1E-30$ (Table 2). Majority of the homoplastic function EST-SSR markers were derived from wound-induced proteins, expressed proteins, and cysteine proteinase inhibitors.

## Diversity and Genetic Relationship Analysis

To examine the genetic variation, POPGENE software version 1.31 was used. Genetic diversity/inter-relationships among seven DNA pools were determined by EST-SSR allelic data analysis. A total of 176 alleles were detected in the 41 polymorphic marker loci. The alleles contained 2–15 SSRs with an average of 4.3 alleles per locus. A total of 30 polymorphic loci with only one allele were selected to perform the preliminary potential biodiversity and cluster analyses. The 30 primer pairs used in this study were highly

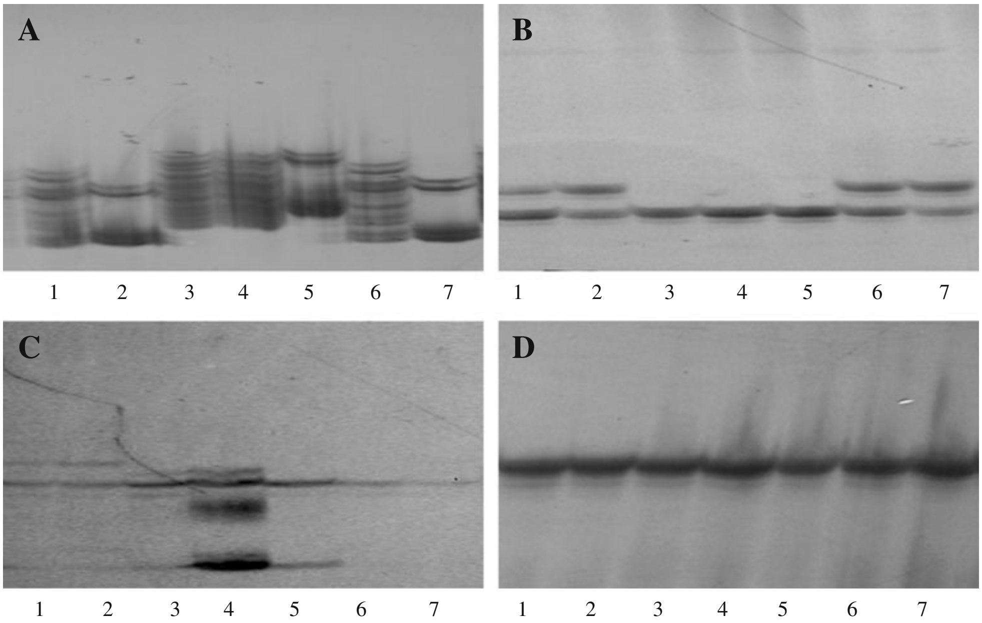

**Fig. 2** Gel showing PCR amplification results using EST-SSR primer pairs. **a** Nonsense primers (70); **b** polymorphic primers (1,656); **c** low quality primers (34); **d** no polymorphism control primers (417). Seven DNA samples from walnuts cultivated in different regions were tested. *Lane 1*, Yecheng County; *2*, *P. hupehensis* Skan; *3*, Tibet walnut; *4*, Hutian County; *5*, Shandong Province; *6*, Wensu County; *7*, Shannxi Province

polymorphic. The number of alleles per locus ranged from one to five, with an average of 2.13. The observed heterozygosity ranged from 0 to 1 (mean Ho=0.457), and the expected heterozygosity in each race ranged from 0.14 to 0.75 (mean He=0.52). An UPGMA tree was constructed after phenetic clustering of these seven DNA pools based on the genotypic data of polymorphic markers. The dendrogram of different DNA pools was obtained based on Nei's (1978) unbiased genetic distance. Obvious relationships between genetic and geographic distance for all detected samples were observed. Samples from Yecheng and Hutian were clustered into one sister group, which had a minor relationship with Wensu samples, as shown in Fig. 3. This is presumably due to the neighborly geographical location of the three counties in Xinjiang province, the oldest walnut cultivation region in China. Meanwhile, samples from Tibet formed an independent group, and the outgroup of wingnut was located at the furthest distance of the dendrogram.

## Discussion

### Evaluation of EST-SSR Frequency in the Walnut Sequence Database

Results revealed that walnut ESTs are rich in microsatellites. A total of 398 SSRs were found in 2,414 unigenes sequences. The average density of SSR distribution was one SSR per 4.18 kb, which was much higher than that of other plants, such as wheat (1/15.6 kb; Kantety et al. 2002), barley (1/6.3 kb; Thiel et al. 2003), Arabidopsis (1/13.83 kb), tomato (1/11.1 kb), cotton (1/20.0 kb), soyabean (1/7.4 kb), poplar (1/14.0 kb; Cardle et al. 2000), and citrus (1/5.7 kb; Chen et al. 2006).

Abundance of SSRs in unigenes varied greatly in different plant species based on the minimum length criteria of the SSR repeat. For instance, EST data mining of rice revealed that the SSR frequency decreased from 50% to 1% when SSR length increased from 12 bp to 30 bp. When the least SSR length was set to 40 bp, the frequency of di- was higher than that of tri-repeats (Rota et al. 2005; Varshney et al. 2005). Similar results were observed in walnut EST data mining. Only 59 SSR loci were found when the mining criteria were set to ten SSR motif repeats. Therefore, we conclude that the overall frequency of detectable SSRs with different length is dependent on the criteria set during data mining, the size of the dataset, and the database mining tools, which was in good accordance with results reported by Varshney et al. (2005).

The abundance of different SSR repeat types also varied across genomes. Data mining of walnut EST-SSRs revealed apparent preferences of certain repeat types in its genome. Results also showed that tri- and di-repeats were the fundamental EST-SSR motifs, and the di-motifs were more prevalent in walnut. Similar results have been reported in spruce (Bérubé et al. 2007) and rape (Li et al. 2007). However, this is contrary to other reports that tri-repeats were the most abundant motifs recovered from plant ESTs (Cardle et al. 2000; Gao et al. 2003; Kantety et al. 2002; Temnykh et al. 2001; Varshney et al. 2002; Tóth et al. 2000). SSR types generally affect the coding of proteins, with the exception of tri-SSRs in ESTs which would not result in frameshift mutations in the coding regions (Varshney et al. 2005). Our result showed that 28.07% of those sequences containing di-repeats were localized in ORFs when com-

**Table 2** Annotation of polymorphic markers and their homologies to protein-coding genes

| Probe | Annotation | Species | *E* value | SSR location |
|---|---|---|---|---|
| Contig_5 | Putative uncharacterized protein | *Vitis vinifera* | 3.00E−25 | ORF |
| Contig_40 | Wound-induced protein | *Mesembryanthemum crystallinum* | 2.00E−30 | 3 |
| Contig_62 | Albumin | *Bertholletia excelsa* | 1.00E−26 | 3 |
| Contig_104 | Putative uncharacterized protein | *Zea mays* | 8.00E−86 | 5 |
| Contig_156 | No hits found | | | |
| Contig_216 | No hits found | | | |
| Contig_259 | Putative uncharacterized protein | *Arabidopsis thaliana* | 1.00E−105 | ORF |
| Contig_268 | Putative uncharacterized protein | *Arabidopsis thaliana* | 1.00E−20 | Unknown |
| Contig_269 | No hits found | | | |
| Contig_352 | MtN3-like protein | *Arabidopsis thaliana* | 1.00E−80 | 3 |
| Contig_407 | No hits found | | | |
| Contig_541 | No hits found | | | |
| Contig_562-2 | Dehydrin (fragment) | *Betula pubescens* | 3.00E−17 | ORF |
| Contig_566 | Putative uncharacterized protein | *Arabidopsis thaliana* | 5.00E−07 | 3 |
| Contig_610 | Wound-induced basic protein | *Phaseolus* | 7.00E−19 | 5 |
| Contig_642 | No hits found | | | |
| Contig_718 | Putative transcription factor | *Arabidopsis thaliana* | 2.00E-61 | 5 |
| Contig_719 | Lysophospholipase-like protein | *Arabidopsis thaliana* | 1.00E−143 | 3 |
| Contig_721 | Expressed protein | *Arabidopsis thaliana* | 3.00E−20 | ORF |
| Contig_795 | Oleosin | *Ficus awkeotsang* | 4.00E−39 | 3 |
| Contig_870 | NifU-like protein | *Arabidopsis thaliana* | 6.00E−64 | 5 |
| Contig_1000 | No hits found | | | |
| Contig_1172 | No hits found | | | |
| Contig_1396 | No hits found | | | |
| Contig_1458 | No hits found | | | |
| Contig_1464 | Putative uncharacterized protein | *Zea mays* | 8.00E−19 | ORF |
| Contig_1483 | Cysteine proteinase inhibitor | *Zea mays* | 8.00E−25 | 3 |
| Contig_1528 | Serine/threonine protein | *Arabidopsis thaliana* | 3.00E−80 | 3 |
| Contig_1529 | No hits found | | | |
| Contig_1552 | Os08g0533600 protein | *Oryza sativa* subsp. | 4.00E−32 | 3 |
| Contig_1625 | Nucleolin | *Zea mays* | 2.00E−30 | ORF |
| Contig_1626 | Ara4-interacting protein | *Arabidopsis thaliana* | 3.00E−32 | ORF |
| Contig_1631 | GT-2 factor (Fragment) | *Soybean* | 2.00E−33 | Unknown |
| Contig_1632 | No hits found | | | |
| Contig_1656 | Putative uncharacterized protein | *Zea mays* | 9.00E−54 | 3 |
| Contig_1681 | Expressed protein | *Arabidopsis thaliana* | 7.00E−85 | 3 |
| Contig_1682 | Senescence-associated protein | *Nicotiana tabacum* | 8.00E−51 | 3 |
| Contig_1687 | No hits found | | | |
| Contig_1692 | No hits found | | | |
| Contig_1693 | DNA-binding protein | *Daucus carota* | 2.00E−39 | Unknown |
| Contig_1712 | Putative uncharacterized protein | *Arabidopsis thaliana* | 8.00E−33 | Unknown |

pared with known protein-coding genes. The high abundance is unexpected since di-repeats would potentially perturb the reading frame of a gene when present in an exon, the region that directly relate to functions of the given species (Bérubé et al. 2007). The high frequency of di-repeats is possibly responsible for encoding strings of amino acids that were prevalent in walnut. Similar explanations have been proposed in several other species when high abundance of dinucleotide repeat motifs of AG was found (Kantety et al. 2002; Bérubé et al. 2007).

```
                                                                      +--------pop1
                                                                +----1
                                                            +--2     +--------pop4
                                                            !   !
                                    +-----------------3   +------------pop6
                                    !                      !
   +------------------------5                     +--------------pop3
   !                                !
   !                                !                       +--------------pop5
--6                                +-----------------4
   !                                                         +-------------pop7
   !
   +--------------------------------------------------------pop2
```

**Fig. 3** UPGMA tree of the seven DNA samples constructed based on 30 EST-SSR markers. The seven DNA pools were *pop1* of Yecheng County; *pop2* of *P. hupehensis* Skan; *pop3* of Tibet walnut; *pop4* of Hutian County; *pop5* of Shandong Province; *pop6* of Wensu County; and *pop 7* of Shannxi Province

## Development and Evaluation of EST-SSR Markers in Walnut

Until now, most of the published SSR primer pairs in walnut were genome-based markers. In our study, a total of 123 EST-SSR markers were identified, providing a significant supplement to the presently available microsatellite markers for walnut research. Considering the least polymorphism being SSR ≤ 12 bp (Temnykh et al. 2001), four SSR loci (<12 bp) primer pairs were initially designed and tested. However, results showed either poor polymorphism or no amplification. Therefore, we subsequently ignored other SSR loci of less than 12 bp. Among the 123 primer pairs, only 41 generated polymorphism by PCR. The proportion is lower than that of Chinese cabbage (46.7%; Xin et al. 2006), barley (46.0%; Thiel et al. 2003), and

**Table 3** Genetic diversity statistics of walnut

| Probe | na | ne | I | Nei’ | Exp_Het | Obs_Het | Fst |
|---|---|---|---|---|---|---|---|
| Contig_5 | 3 | 1.5556 | 0.656 | 0.3571 | 0.3846 | 0.4286 | 0.4 |
| Contig_62 | 2 | 1.96 | 0.6829 | 0.4898 | 0.5275 | 0.8571 | 0.125 |
| Contig_156 | 4 | 2.9697 | 1.1973 | 0.6633 | 0.7143 | 0.8571 | 0.3538 |
| Contig_259 | 3 | 2.2791 | 0.8982 | 0.5612 | 0.6044 | 1 | 0.1091 |
| Contig_268 | 2 | 1.6897 | 0.5983 | 0.4082 | 0.4396 | 0.5714 | 0.3 |
| Contig_269 | 2 | 1.3243 | 0.4101 | 0.2449 | 0.2637 | 0 | 1 |
| Contig_407 | 2 | 1.3243 | 0.4101 | 0.2449 | 0.2637 | 0.2857 | 0.4167 |
| Contig_610 | 2 | 1.1529 | 0.2573 | 0.1327 | 0.1429 | 0.1429 | 0.4615 |
| Contig_642 | 3 | 2.3333 | 0.9557 | 0.5714 | 0.6154 | 0.2857 | 0.75 |
| Contig_718 | 4 | 2.8 | 1.171 | 0.6429 | 0.6923 | 0.2857 | 0.7778 |
| Contig_719 | 3 | 1.5556 | 0.656 | 0.3571 | 0.3846 | 0.2857 | 0.6 |
| Contig_721 | 4 | 2.3902 | 1.0547 | 0.5816 | 0.6264 | 0.8571 | 0.2632 |
| Contig_795 | 2 | 1.3243 | 0.4101 | 0.2449 | 0.2637 | 0 | 1 |
| Contig_870 | 3 | 2.3333 | 0.9557 | 0.5714 | 0.6154 | 0 | 1 |
| Contig_1000 | 4 | 3.3793 | 1.2721 | 0.7041 | 0.7582 | 0.1429 | 0.8986 |
| Contig_1172 | 3 | 2.6486 | 1.0346 | 0.6224 | 0.6703 | 0.5714 | 0.5410 |
| Contig_1396 | 3 | 2.5128 | 0.9923 | 0.602 | 0.6484 | 1 | 0.1695 |
| Contig_1458 | 3 | 2.8 | 1.0609 | 0.6429 | 0.6923 | 0.8571 | 0.3333 |
| Contig_1464 | 4 | 1.8491 | 0.8953 | 0.4592 | 0.4945 | 0.5714 | 0.3778 |
| Contig_1483 | 2 | 1.5077 | 0.5196 | 0.3367 | 0.3626 | 0.1429 | 0.7879 |
| Contig_1528 | 4 | 3.1613 | 1.2397 | 0.6837 | 0.7363 | 1 | 0.2687 |
| Contig_1529 | 2 | 1.96 | 0.6829 | 0.4898 | 0.5275 | 0 | 1.0000 |
| Contig_1625 | 5 | 3.0625 | 1.2964 | 0.6735 | 0.7253 | 1 | 0.2567 |
| Contig_1626 | 2 | 1.96 | 0.6829 | 0.4898 | 0.5275 | 0 | 1 |
| Contig_1631 | 3 | 1.5556 | 0.656 | 0.3571 | 0.3846 | 0.1429 | 0.800 |
| Contig_1656 | 4 | 2.1778 | 1.0285 | 0.5408 | 0.5824 | 0.5714 | 0.4717 |
| Contig_1687 | 3 | 1.5556 | 0.656 | 0.3571 | 0.3846 | 0.4286 | 0.4000 |
| Contig_1692 | 2 | 1.6897 | 0.5983 | 0.4082 | 0.4396 | 0 | 1.000 |
| Contig_1693 | 4 | 3.1613 | 1.2397 | 0.6837 | 0.7363 | 1 | 0.2687 |
| Contig_1712 | 3 | 1.8148 | 0.7963 | 0.449 | 0.4835 | 0.4286 | 0.5227 |

Only the 30 functional EST-SSR markers which amplified one allele were included in the analysis

*na* observed number of alleles, *ne* effective number of alleles, *I* Shannon's Information index, *Exp_Het* expected heterozygosity, *Obs_Het* observed heterozygosity, *Nei'* Nei’s expected heterozygosity, *Fst* fixation index

*Camellia sinensis* (62.5%; Jin et al. 2006). The remaining 82 primer pairs were excluded for further analysis. The indetermination of designed EST-SSR primer has been reported elsewhere, and functional primer rates ranged from 60% to 90% in genomic and EST-SSRs, respectively (Thiel et al. 2003; Kota et al. 2001; Yu et al. 2004; Saha et al. 2004; Cordeiro et al. 2001; Gupta et al. 2003). Possible explanations include the use of questionable sequence information for primer development, one or both of the EST-SSR primer pair extending across a splice site, presence of large introns in genomic DNA sequence, and primer pairs being derived from chimeric cDNA clones (Varshney et al. 2005). Thus, the quality of the SSR-EST sequences is important for primer design. Thiel et al. (2003) and Sreenivasulu et al. (2002) found that up to 9% of cereal ESTs were of low quality and should be rejected for primer designing. Meanwhile, as Thiel et al. (2003) reported, even though the sequence data of barley was of high quality, rigorous quality check of the EST sequences and careful clustering were still essential for minimizing the failures in EST-SSR marker development. In addition, some other strategies were also reported to optimize primer design, including avoiding PCR products larger than 700 bp, choosing one single locus sequence as the ideal motif and minimizing fussy parameters (Varshney et al. 2005; Thiel et al. 2003).

In the present investigation, 16 primer pairs were detected with no polymorphic amplification among the tested walnut accessions. Smulders et al. (1997) ascribed it to the initial stage of mutational decay with less chance for polymerase slippage. Additionally, the observed polymorphism was occasionally caused by a size polymorphism with the intron, which may overshadow a putative polymorphism of the EST-based microsatellite. Fragments over 500 bp cannot be scored accurately for small differences in fragment size. Such amplification problems caused by introns have been reported by Thiel et al. (2003) in barley.

As EST-SSRs are present in gene-rich regions of the genome, they exhibit less polymorphism than genomic-based SSRs. In walnut, putative functions were matched to 34.1% of the polymorphic SSR-ESTs markers out of the 41 polymorphic SSR-EST datasets using a stringent criterion ($E$ value≤1E−30). Our results showed that 14 SSR-ESTs had no hits in the protein database (34.1%), and 19.5% were annotated as putative uncharacterized proteins. These putative proteins might represent some specific transcripts of walnut with unknown functions. After comparison with known proteins in silico, SSR markers showing polymorphism were found to mainly locate in the 3′-untranslated region. We found only two polymorphic SSR loci located in the ORF. Although heavy selection against frameshift mutations existed in plant kingdom, there is no doubt that the SSRs do exist in ORFs, which could potentially disturb the reading frame of exons and result in the change of amino acid sequences and subsequently protein functions.

SSRs were believed to be locus specific. Therefore, one generally expects to amplify a single band or co-migrating twin bands with a single SSR primer set. However, in our study, a total of 11 primer pairs generated multiple bands, which may due to the amplification of more than one homo-locus by each EST-SSR (Holton et al. 2002). Those SSR loci corresponding to one allele were analyzed subsequently, and 30 EST-SSR markers were used in the genetic relationship test (Table 3). The association between genetic similarity and geographic proximity was indicated by Nei's genetic distance (Fig. 3) and pairwise Fst values. The accessions from Yecheng, Hutian, and Wensu counties were clustered together. These three places are geographically close in Xinjiang Province, the presumed original area of walnut cultivation in China. The close relationship among these accessions is probably due to (1) the natives selected and introduced the elite cultivars from one place to another in history; (2) in ancient times, the walnut trees and seeds dispersed via trade caravan routes because all three counties were located on the ancient “Silk Road.” Same dispersal result has also been observed in apricot, *Coriandrum sativum*, etc., which originated from Xinjiang province (He et al. 2006). Meanwhile, accessions from Tibet Province were clustered as an independent group, in agreement with the previous morphology studies of the distinct ecotype distributions in China (Xi 1987). Moreover, results from this study provided some molecular evidence to our hypothesis that relatively segregative geographic environment contributes to the formation of the Tibet walnut geoecotype.

## Conclusion

This study demonstrated that data mining of microsatellite loci from database sequences was a simple yet efficient means for EST-SSR marker development. The high abundance of microsatellites in transcribed regions of a genome, and properly designed polymorphic markers, will make EST libraries valuable resources for downstream molecular studies, such as genetic mapping of walnut.

**Acknowledgments** We thank Drs. Chunxian Chen (Citrus Research and Education Center, University of Florida), Zhaocheng Ma, and Qiang Xu (National Key Laboratory of Crop Genetic Improvement) for critically reviewing the manuscript. This research was supported by the doctoral fund of Xinjiang Production and Construction Corp. (2006JC06).

## References

Aggarwal RK, Hendre PDS, Varshney RVK, 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

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, Young AC, Huang S, Gmitter FG Jr (2006) Mining and characterizing microsatellites from citrus ESTs. Theor Appl Genet 112:1248–1257

Cheng YJ, Guo WW, Yi HL, Pang XM, Deng XX (2003) An efficient protocol for genomic DNA extraction from Citrus species. Plant Mol Biol Report Plant Mol 21:177a–177g

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

Dangl GS, Woeste K, Aradhya MK, Koehmstedt A, Simon C (2005) Characterization of 14 microsatellite markers for genetic analysis and cultivar identification of walnut. J Am Soc Hortic Sci 130 (3):348–354

Eujayl I, Sledge MK, Wang L, May GD, Chekhovskiy K, Wonitzer JCZ, Mian MAR (2004) *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago spp*. Theor Appl Genet 108:414–422

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

Gao LF, Tang JF, Li HW, Jia JZ (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. J Mol Breed 12:245–261

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomics 270:315–323

Hackauf B, Wehling P (2002) Identification of microsatellite polymorphisms in an expressed portion of the rye genome. Plant Breeding 121:17–25

He TM, Chen XS, Gao JS, Zhang DH, Xu LI, Wu Y (2006) Using SSR markers to study population genetic structure of cultivated apricots native to Xinjiang. Acta Hortic Sinica 33(4):809–812

Holton TA, Christopher JT, McClure L, Harker N, Henry RJ (2002) Identification and mapping of polymorphic SSR markers from expressed gene sequences of barley and wheat. Mol Breed 9:63–71

Jin JQ, Cui HR, Chen WY, Lu MZ, Yao YL, Xin Y, Gong XC (2006) Data mining for SSRs in ESTs and development of EST-SSR marker in tea plant (Camellia sinensis). J Teach Sci 26(1):17–23

Kantety RV, Rota ML, 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

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

Li XB, Zh ML, Cui HR (2007) Data mining for SSRs in ESTs and development of EST-SSR marker in oilseed rape. J Mol Cell Biol 40:138–144

Li Q (2003) Downloaded from http://file.biopatent.cn/3533

Nei M (1973) Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences, USA 70:3321–3323

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–590

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

Rota ML, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat and barley. J BMC Genomics 6:1–12

Saha MC, Mian MAR, Eujay I, Zwonitzer JC, Wang LJ, May GD (2004) Tall fescue EST-SSR markers with transferability across several grass species. Theor Appl Genet 109:783–791

Smulders MJM, Bredemeijer G, Arens WRP, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among Lycopersicon esculentum cultivars and accessions of other Lycopersicon species. Theor Appl Genet 97:264–272

Sreenivasulu N, Kishor PBK, Varshney RK, Altschmied L (2002) Mining functional information from cereal genomes: the utility of expressed sequence tags. Curr Sci 83:965–973

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.). 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

Tóth G, Gáspári 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

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

Xi Shk (1987) Gene resources of Juglans and genetic improvement of Juglans regia in China. Scientia Silvae Sinicae 31:342–349, China

Xin Y, Cui HR, Lu MZ, Yao YL, Jin JQ, Lim Y, Choi SY (2006) Data mining for SSRs in ESTs and EST-SSR marker development in Chinese cabbage. Acta Horticult Sinica 03:549–554

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

Yeh FC, Boyle TJB (1997) Population genetic analysis of codominant and dominant markers and quantitative traits. Belg J Bot 129:157

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