TÜBİTAK

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2015) 39: 439-450

doi:10.3906/tar-1406-144

# Identification and functional annotation of expressed sequence tags based SSR markers of *Stevia rebaudiana*

**Rajinder KAUR[1,*], Neha SHARMA[2], Ravinder RAINA[3]**

[1]Department of Biotechnology, Dr Y S Parmar University of Horticulture and Forestry, Nauni, Solan (HP), India
[2]Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, Punjab, India
[3]Department of Forest Products, Dr Y S Parmar University of Horticulture and Forestry, Nauni, Solan (HP), India

**Received:** 25.06.2014 • **Accepted:** 08.01.2015 • **Published Online:** 12.06.2015 • **Printed:** 30.06.2015

**Abstract:** *Stevia rebaudiana* Bertoni, a perennial herb belonging to the family Asteraceae, is a tropical as well as subtropical plant of medicinal importance. Despite being an important medicinal plant and in fact proving a boon to diabetics, it has remained almost ignored by molecular biologists. A systematic molecular evaluation of stevia germplasm and molecular breeding work are sparsely reported mainly due to the limited specific molecular markers available. Expressed sequence tags (ESTs) are a valuable resource for developing simple sequence repeats (SSRs). In this study, 5548 stevia ESTs sequences from leaf tissues were retrieved from the NCBI database. Clustering and assembly of these ESTs resulted in a nonredundant set of sequences, i.e. a total of 471 contigs and 3845 singleton ESTs. Out of these 471 contigs and 3845 singletons, 168 SSRs were identified. Dinucleotide SSR is the dominant repeat type (62.5%), followed by trinucleotide (37.5%). Annotation analysis revealed that putative function could be assigned to 82.2% of EST-SSRs. After using Primer3 software, 18 primers were custom synthesized from the SSR containing 18 singletons and generated 61.11% polymorphism. In silico mined markers with functional annotations, in the present study, facilitated genome analysis in stevia, which had not been performed previously. Additionally, the EST-SSRs will be used for molecular work in related plant species since they generally exhibit cross species transferability, making further work cost effective and simple.

**Key words:** EST, functional annotation, genetic diversity, SSR, *Stevia rebaudiana*

## 1. Introduction

Expressed sequence tags (ESTs) are created by partially sequencing randomly isolated gene transcripts that have been converted into cDNA (Adams et al., 1992). Any actively growing tissue can be used to generate cDNAs from mRNA and sequenced to generate ESTs that are assembled into a nonredundant set of sequences (contigs and singletons). Recent research has revealed that ESTs are a potentially rich source of SSRs that reveal polymorphism.

SSRs repeat motifs of 1–6 bp in length are hypervariable and widely spread in eukaryotic genomes (Vosman and Arens, 1997; Rallo et al., 2000). SSRs are reproducible, multiallelic, co-dominant, abundant, and cover the whole genome and have become markers of choice for diversity analysis and genome analysis (Morand et al., 2002; Zeid et al., 2003; Yi et al., 2006; Jiao et al., 2012; Wijarat et al., 2012).

However, the traditional genomic library dependent approach for SSR marker development is time consuming and cannot be carried out in a laboratory without much of an infrastructure. Over the last 10–15 years there has been continuous progress on EST sequencing work in different laboratories that has led to in silico data mining for ESTs and their applications in various molecular biology projects. With the advances in bioinformatics, it is possible to mine and analyze large-scale EST datasets efficiently and exhaustively in various organisms (Ewing et al., 1999; Ronning et al., 2003). An in silico database offers the opportunity to identify SSRs in ESTs through data mining and provides a simple and cost efficient approach for the development of SSR markers in plants (Gupta and Varshney, 2000). Currently there are near 20 million ESTs in the NCBI public collection-dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/).

EST-SSRs have been reported in several plant species, such as cauliflower (Vaidya et al., 2012), walnut (Zhu et al., 2009), watermelon (Verma and Arya, 2008), citrus (Chen et al., 2006), sugarcane (Pinto et al., 2004), grapes (Scott et al., 2000), cereal species (Varshney et al., 2002), and bread wheat (Gupta et al., 2003). EST-derived markers have also been well documented in marker assisted selection that involves mainly the identification of QTL of important

* Correspondence: rkaur_uhf@rediffmail.com

traits, genetic mapping, genetic diversity, and comparative genomics (Liu et al., 2005; Alwala et al., 2006; Miklas et al., 2006; Cloutier et al., 2009; Durand et al., 2010; Kumar et al., 2011; Vaidya et al., 2012).

*Stevia rebaudiana* is an important medicinal plant with antibacterial, antifungal, antiinflammatory, antiviral, antiyeast, cardiotonic, and diuretic properties. The property of this species that has attracted attention is the intense sweet taste of its leaves and aqueous extracts. Basically, stevia is distinguished by the presence of sweet diterpene glycosides: rebaudioside- A, rebaudioside-C, stevioside, and dulcoside in its leaf tissue (Ahmed and Dobberstein, 1982). A total of nine sweet diterpenoid glycosides have been isolated from *S. rebaudiana*, namely: stevioside, steviolbioside, dulcoside-A, and rebaudioside-A, -B, -C, -D, -E, and -F (Kennelly, 2002; Starratt et al., 2002; Dacome et al., 2005). Stevioside is the major component but it has an unpleasant bitter aftertaste (Ishima and Katayama, 1976; Kinghorn and Soejarto, 1985). However, rebaudioside-A, normally present in lower amounts (25% to 45% of stevioside content) in leaves, does not have bitter aftertaste and has a sweetening power 1.2 to 1.6 times higher than that of stevioside. Therefore, the breeding objectives can be set for stevioside and rebaudioside-A content for this medicinally important plant. Earlier, Heikal et al. (2008) employed ISSR markers to determine the phylogenetic relationships among six stevia collections and Yao et al. (1999) first constructed its genetic linkage map. However, the genetic studies on stevia are chronically lagging behind compared with other economically important plant species.

To date, publically available molecular markers such as SSRs in stevia are limited, which has significantly hampered genome analysis and molecular breeding work. Therefore, the present study was undertaken with the following objectives: 1) Identification of SSRs in the transcribed regions of the *Stevia* genome; 2) Function domain analysis. The present study is an initiative in the direction of development of functional stevia-specific microsatellite markers and the main aim of the study was for these EST-SSR markers to be efficiently used for genetic mapping studies of segregating stevia populations.

## 2. Materials and methods

### 2.1. Source plant material

A survey was conducted to collect diverse *Stevia rebaudiana* plants from all 12 districts of Himachal Pradesh (a state in Northern India). During the survey, it was found that stevia is present in a few pockets of only four districts of the state. The plants were collected from these four districts, viz., Kangra (Palampur, Institute of Himalayan Bioresource Technology (IHBT) and Dharamshala, from a local farmer) at an altitude of 733 m above mean sea level (amsl), Hamirpur (Hamirpur, near Institute of Biotechnology and Environmental Science (Neri) and Herbal Garden of Beri) at an altitude of 738 m amsl, Bilaspur (from a local farmer) at an altitude of 673 m amsl, and Solan (Nauni, Dr Y S Parmar University of Horticulture and Forestry, and Chail) at an altitude of 1275 m amsl.

Additionally, stevia plants were collected from the Indian Agricultural Research Institute (IARI), New Delhi, at an altitude of 250 m amsl. All the collections used for the present study are cited in Table 1.

### 2.2. EST search by *Stevia* EST mining

EST sequences belonging to *Stevia rebaudiana* leaf cDNA were acquired from NCBI in their FASTA format from the EST database (www.ncbi.nlm.nih.gov/nucest). A total 5548 ESTs were retrieved and kept as a single text file constructed containing all the detected sequences.

### 2.3. Clustering and assembly of EST sequences

A total of 5548 redundant ESTs retrieved from NCBI were used to produce a nonredundant dataset by using the EGassembler webserver (Masoudi-Nejad et al., 2006). The repetitive elements including small RNA pseudo genes, LINEs, SINEs, LTR elements, vector sequences, organelle, and other interspersed repeats were masked by the software. The software automatically screens and cleans for various contaminants in the retrieved EST sequences. The server clustered and assembled the sequences into contigs and singletons using CAP3 (Huang and Madan, 1999) with the criterion of 80% overlap identity between one end of a default read to another end.

### 2.4. SSR identification for the assembled ESTs

The potential SSRs were detected in the assembled ESTs by submitting the sequences to a search tool, namely SSRIT (simple sequence repeat identification tool; www.gramene.org/db/searches/SSRtool) (Temnykh et al., 2001). The sequence search for SSRs conducted by SSRIT was carried out by setting the search parameters to identify at least five repeat units SSR of maximum ten base pairs. The software needs to put input as a FASTA format sequence in which SSRs are to be found and the output file comprises the sequence name in which the SSR is detected, repeat motif of the SSR, number of times it is repeated, start and end of the repeat, and length of the sequence. The SSR motifs obtained as output of SSRIT were thoroughly counted for the number of repeats and their frequency and distribution were also recorded.

### 2.5. EST-SSR similarity search or BLASTx analysis

The frequency and nature of amino acids generated from the trinucleotide repeat motifs were studied because each trinucleotide motif codes an amino acid that has putative roles in the biological activity of protein molecules. To find out the putative functions of the EST SSRs, the

**Table 1.** List of 16 *Stevia rebaudiana* collections used in the present study for EST-SSR analysis.

| S. no. | Denoted as | Place of collection | District |
|---|---|---|---|
| 1. | Parent A | Nauni, Y S Parmar UHF | Solan |
| 2. | Parent B | Nauni, Y S Parmar UHF | Solan |
| 3. | St-1 | Palampur, IHBT | Kangra |
| 4. | St-2 | Dharamshala, from a local farmer | Kangra |
| 5. | St-3 | Neri, Research station | Hamirpur |
| 6. | St-4 | Beri, Herbal Garden | Hamirpur |
| 7. | St-5 | Bilaspur, from a local farmer | Bilaspur |
| 8. | St-6 | New Delhi, IARI | New Delhi |
| 9. | Nst-1 | Nauni, Y S Parmar UHF | Solan |
| 10. | Nst-2 | Nauni, Y S Parmar UHF | Solan |
| 11. | Nst-3 | Nauni, Y S Parmar UHF | Solan |
| 12. | Cst-1 | Chail | Solan |
| 13. | Rst-1 | Replication cutting of St-1 | Kangra |
| 14. | Rst-2 | Replication cutting of St-2 | Kangra |
| 15. | Rst-5 | Replication cutting of St-5 | Bilaspur |
| 16. | Rst-6 | Replication cutting of St-6 | New Delhi |

corresponding EST sequences were compared with the UniProt database (http://www.uniprot.org/) using the BLASTx program, assuming an E value of ≤1E – 5 as a significant criterion of homology.

**2.6. Functional domain analysis**

Using PROSITE from ExPASy tools (http:/prosite.expasy.org/prosite.html) functional protein domains of the sequences containing SSRs were searched.

**2.7. Primer design and polymerase chain reaction (PCR) analysis**

Using the Primer3 software (www.frodo.wi/mit.edu/primer3) 18 primers were custom synthesized from the SSR containing 18 singletons randomly chosen from a total of 3845 singletons and these primers were used for the amplification of isolated DNA of collections of *Stevia rebaudiana*.

The PCR protocol was standardized to amplify the genomic DNA of stevia plants using EST-SSR primers. PCR was carried out in the presence of 60 ng of genomic DNA, 1X PCR buffer, 2.2 mM $MgCl_2$, 1 mM dNTPs, 3.0 µM of each primer, and 1 Unit 94 kD Taq DNA polymerase (all chemicals from Genei, Bangalore, India). The PCR amplification was performed in an automated thermal cycler (Applied Biosystems, USA) programmed for 5 min initial denaturation at 95 °C, 40 cycles of 1 min denaturation at 94 °C, 1 min annealing according to Tm for each primer and 2 min extension at 72 °C for 2 min, followed by a final extension at 72 °C for 5 min. The PCR amplified products were subjected to electrophoresis in a 2% agarose gel at a constant voltage of 80 V. After electrophoresis, the gels were stained with 0.5 µg/mL ethidium bromide and a photograph taken under UV light in Gel Documentation system (Syngene, USA).

**2.8. Data analysis**

Two different programs, NTSYSpc ver. 2.02h (Rohlf, 2000) and DARwin5 ver. 5.0.155 (http://darwin.cirad.fr/darwin) (Perrier et al., 2003), were used to study the relatedness of the diverse *Stevia* plants. The data on band position on the gel were recorded by assigning '0' for the absence of band and '1' for the presence of band. After that, a dendrogram, similarity matrix tables, and rooted trees were constructed by using the UPGMA function of both programs.

Genetic diversity, defined as the polymorphism information content (PIC; Anderson et al., 1993), was used to measure allelic diversity at each locus. PIC values were calculated as follows:

$PIC = 1 - \Sigma^{k}_{i=1} pi^2$,

where k was the total number of alleles detected for a microsatellite (SSR) marker and *pi* was the frequency of the ith allele in the set of genotypes analyzed.

## 3. Results

**3.1. NCBI database search**

A total of 5548 EST sequences belonging to stevia were obtained from NCBI and used for further computations and analysis.

Out of the 5548 ESTs, 471 contigs were assembled and 3845 singletons were recorded that showed no overlap with any ESTs. After assembly, the whole dataset was reduced to 4316 sequences, presented in Table 2, which showed 22.2% data redundancy by using EGassembler. These nonredundant sequences were reported to have 168 EST-SSRs in which 28 EST-SSRs were reported by contigs and remaining 140 were from singletons. After assembly, a nonredundant group of ESTs became nonredundant, which included contigs and singletons, and these nonredundant ESTs were used for the SSR search tool.

### 3.2. SSRIT search

The SSRIT search for microsatellites or SSRs detected a total of 168 SSRs. Analysis of the detected SSRs revealed that all of them represented di- and trinucleotide repeats. From Table 3 it is seen that dinucleotide SSR is the dominant repeat type (62.5%), followed by trinucleotide (37.5%). Among the trinucleotide repeat motifs the dominant repeat motif is ACC (12.7%), followed by ATC (11.1%). Table 4 clearly depicts that among dinucleotide repeats the most frequent repeat was CT/TC (32.3% followed by AT/TA (27.6%), while the least frequent dinucleotide was GC/CG (3.81%).

### 3.3. Frequency and nature of amino acids generated from the trinucleotide repeat motifs

With the advent of bioinformatics, proteins have become increasingly studied at a genomic level. As a result, the practice of representing the genetic code as a DNA codon has become more popular. In the present study, each trinucleotide motif codes an amino acid that has putative roles in the biological activity of protein molecules. Table 5 clearly revealed that out of a total of 63 trinucleotides 11.11% of trinucleotides SSRs encoded isoleucine, which is nonpolar, and in total 47.06% polar and 52.94% nonpolar amino acids were detected.

### 3.4. BLASTx analysis and functional domain analysis

A total of 163 sequences were judged for BLASTx analysis and, based on this analysis, a putative function was assigned to 134 of the sequences (82.2%). The annotation results indicated that most of the EST-SSR sequences showed high homology with known *Arabidopsis thaliana*, *Helianthus annuus*, and *Medicago truncatula* proteins. Functional domain analysis revealed that most of the EST-SSR sequences found the protein domains responsible for VWFC domain signature, EGF-like domain signature 1, and 2Fe-2S ferredoxin-type iron-sulfur binding region signature. These regions are generally important for the function of a protein and/or for the maintenance of its three-dimensional structure. Based on the result of the PROSITE ExPASy tool, out of 163, 150 sequences (92.02%) containing protein domains were identified, which are presented in Table 6. A total of 11 polymorphic primers were obtained, while six primers were found to be monomorphic and one primer did not show any amplification.

### 3.5. PCR amplification

After using the Primer3 software (www.frodo.wi/mit.edu/primer3) 18 primers were custom synthesized from the SSR containing 18 singletons randomly chosen from a total of 3845 singletons and these tested primers generated 61.11% polymorphism. The results indicated that out of all 18 EST-SSRs, 17 primer pairs gave scorable bands, 11 of which were polymorphic and 6 were monomorphic. One EST-SSR did not give any PCR amplification. The alleles produced were varied from single to multiallelic and the amplicon size varied from 100 bp to 1200 bp only. A total of 241 fragments were generated by all 17 primers that showed amplification, i.e. 14.2 fragments per primer.

Jaccard's similarity matrix coefficient obtained through NTSYSpc is presented in Table 7, generated through EST-SSR markers.

Table 7 clearly shows that EST-SSR primer coefficient values ranged from 0.295 to 0.824. The highest value of similarity (0.824) was found between NSt-1 and NSt-3; both were collected from Dr Y S Parmar University of Horticulture and Forestry, Nauni, Solan District. The minimum correlation coefficient (0.295) was obtained between St-3 collected from Neri, Research station, Hamirpur, and "A". The average correlation coefficient for EST-SSR obtained was 0.56.

According to the data obtained from Table 8 EST-SSR primer dissimilarity coefficient values ranged from 0.111 to 0.477. The highest value of dissimilarity (0.477) was found between St-3 collected from Neri, Research station, Hamirpur, and "A" collected from Dr Y S Parmar University of Horticulture and Forestry, Nauni, Solan District. NSt-1 and NSt-3, both collected from Dr Y S Parmar University of Horticulture and Forestry, Nauni, Solan District, were found to have a minimum dissimilarity value of 0.111. The results obtained were in agreement with the results produced by Jaccard's similarity coefficients.

**Table 2.** Results of EST sequence assembly generated by EGassembler.

| Name of crop | Total Number of EST sequences | Number of singletons | Number of contigs | Nonredundant data set (singletons + contigs) | Reduction in redundancy |
|---|---|---|---|---|---|
| *Stevia* | 5548 | 3845 | 471 | 4316 | 22.2% |

**Table 3.** Summary of SSRs found in EST sequences obtained from NCBI.

| SSR type | Total number | Frequency |
|---|---|---|
| Dinucleotides | 105 | 62.5% |
| Trinucleotides | 63 | 37.5% |
| Total | 168 | |

At 38% similarity "A" was singled out from the rest of the 15 collections. Maximum similarity was observed between NSt-1 and NSt-3, both Solan collections of stevia, collected from Dr Y S Parmar UHF, Nauni; and Nst-2, again collected from Dr Y S Parmar UHF, Nauni, was found to be included in the same cluster. Collection Rst-6 (replication cutting of St-6, collected from IARI, New Delhi) was found to be separate from the rest of the 14 collections as it clustered out in the RAPD and ISSR cluster. In the EST-SSR cluster analysis it was clearly observed that at around 70% similarity St-1, St-2 and Rst-1, Rst-2 formed similar clusters as Rst-1 and Rst-2 were replication cuttings of St-1 and St-2, respectively, collected from Palampur, IHBT, and Dharamshala, from a local farmer, District Kangra (Figure). Polymorphism information content for EST-SSR primers ranged from 0.00 to 0.71 and hence gave an average polymorphism information content of 0.355.

## 4. Discussion

In recent years, ESTs have become a valuable tool for genome analyses and are currently the most widely used approach for structural and functional plant genomes (Pashley et al., 2006). EST analysis provides a simple strategy for studying the transcribed regions of genomes; before detection of SSRs in ESTs, the ESTs must be trimmed to reduce the redundancy.

**Table 4.** Distribution of repeat motifs found in EST sequences obtained from NCBI.

| S. No. | Repeat Motif | Number | Frequency |
|---|---|---|---|
| Trinucleotide repeats | | | |
| | ATC/CAT/TCA | 7 + 4 + 3 = 14 | 22.2% |
| | TAG/ATG/GAT/GTA/TGA | 6 + 2 + 4 + 1 + 1 = 14 | 22.2% |
| | ACC/CCA | 8 + 2 = 10 | 15.87% |
| | GGT/TGG | 1 + 2 = 3 | 4.76% |
| | GAA | 6 | 9.52% |
| | CCG/GCC | 1 + 1 = 2 | 3.17% |
| | AAC | 1 | 1.58% |
| | AAT | 2 | 3.17% |
| | CGG | 1 | 1.58% |
| | AGA | 2 | 3.17% |
| | CTT | 3 | 4.76% |
| | GCA | 1 | 1.58% |
| | TGC | 1 | 1.58% |
| | TAT/TTA | 1 + 1 = 2 | 3.17% |
| | CAC | 1 | 1.58% |
| | Total | 63 | |
| Dinucleotide repeats | | | |
| 1. | GA/AG | 5 + 10 = 15 | 14.30% |
| 2. | AC/CA | 10 + 6 = 16 | 15.23% |
| 3. | CT/TC | 15 + 19 = 34 | 32.40% |
| 4. | TG/GT | 2 + 5 = 7 | 6.66% |
| 5. | AT/TA | 22 + 7 = 29 | 27.60% |
| 6. | GC/CG | 1 + 3 = 4 | 3.81% |
| | Total | 105 | |

**Table 5.** List of amino acids and their occurrence frequency depending upon the trinucleotide repeat motifs obtained.

| S. no. | Amino acid | Trinucleotide | No. of times repeated | Frequency of occurrence | Nature of amino acid |
|---|---|---|---|---|---|
| | Isoleucine | ATC | 7 | 11.11% | Nonpolar |
| | Histidine | CAT, CAC | 4 + 1 = 5 | 7.93% | Polar |
| | Serine | TCA | 3 | 4.76% | Polar |
| | Stop codon | TAG | 6 | 9.52% | - |
| | Methionine | ATG | 2 | 3.18% | Nonpolar |
| | Aspartic acid | GAT | 4 | 6.35% | Polar |
| | Valine | GTA | 1 | 1.59% | Nonpolar |
| | Stop codon | TGA | 1 | 1.59% | - |
| | Threonine | ACC | 8 | 12.7% | Polar |
| | Proline | CCA, CCG | 2 + 1 = 3 | 4.76% | Nonpolar |
| | Glycine | GGT | 1 | 1.59% | Polar |
| | Tryptophan | TGG | 2 | 3.17% | Nonpolar |
| | Glutamic acid | GAA | 6 | 9.53% | Polar |
| | Alanine | GCC, GCA | 1 + 1 = 2 | 3.17% | Nonpolar |
| | Asparagine | AAC, AAT | 1 + 2 = 3 | 4.76% | Polar |
| | Arginine | CGG, AGA | 1 + 2 = 3 | 4.76% | Polar |
| | Leucine | CTT | 3 + 1 = 4 | 6.35% | Nonpolar |
| | Cysteine | TGC | 1 | 1.59% | Nonpolar |
| | Tyrosine | TAT | 1 | 1.59% | Nonpolar |

In stevia information on the existence of molecular markers is limited. A few earlier studies have demonstrated the same number of ESTs in stevia as has been reported in the current study. However, none of the studies have reported the detection of SSRs from stevia ESTs. Ours is the first study to generate EST-SSRs and use them for diversity analysis as well as for functional annotations. ESTs, being transcribed regions of the genome, provide a valuable source for candidate gene discovery.

Brandle et al. (2002) and Richman et al. (2005) studied ESTs in stevia and reported that they can be of great help in identification of candidate genes for secondary metabolites specifically for diterpenes, which forms a part of steviol glycoside.

In recent years, in silico approaches have been effectively utilized to detect SSR amplifying genic regions (Igarashi et al., 2008; Woodhead et al., 2008). The development of SSR markers following the conventional approach of genomic library construction is time consuming (Zane et al., 2002) and resource intensive (Squirrell et al., 2003).

Among all the SSR motifs, dinucleotide SSRs were abundantly present. Among trinucleotide repeat motifs the dominant repeat motif was ACC, followed by ATC. Among dinucleotide repeats the most frequent repeat was CT/TC, followed by AT/TA, and the least frequent dinucleotide was GC/CG. This was found to be in agreement with studies in the traditional Chinese medicinal plant *Epimedium sagittatum* (Zeng et al., 2010) and also in coffee (Aggarwal et al., 2007). Di-tetra or pentanucleotide variation would produce shifts in the reading frame that would result in negative selection and a lower degree of polymorphism (Varshney et al., 2002). Hence, in the present study, trinucleotide repeats in SSRs were given preference. The trinucleotide SSRs are triplet codons that code for a particular amino acid. The triplet codons form an open reading frame (ORF) translated to proteins. Out of a total of 63 trinucleotides, the majority of trinucleotides SSRs encoded isoleucine, which is nonpolar. The analysis of data also revealed that the majority of amino acids were nonpolar.

The annotation results of BLASTx analysis indicated that most of the EST-SSR sequences of *S. rebaudiana* showed high homology with known proteins of *Arabidopsis thaliana*, *Helianthus annuus*, and *Medicago truncatula* proteins, and functional domain analysis revealed that most of the EST-SSR sequences found the protein domains responsible for VWFC domain

**Table 6.** BLASTx, functional domain analysis, singletons that showed homology to genes with proteins matching *Stevia rebaudiana* identified in a BLASTx search of the NCBI database and SSR markers generated.

| S.no. | Sequence Id | Protein | Source | Primer sequence | E-value | Percent similarity | Protein domain | PIC | Polymorphic/ monomorphic and amplicon size range |
|---|---|---|---|---|---|---|---|---|---|
| 1. | gi\|18465673\| | hAT-like transposase | *Arabidopsis thaliana* | F: CGGGTTAGAAGGAAACGTGA<br>R: AAGTTTCCACCAACCCATCA | 9e - 59 | 63% | 2Fe - 2S ferredoxin -type iron - sulfur binding region signature | 0.71 | Polymorphic 500 bp to 800 bp. |
| 2. | gi\|18465467\| | expressed protein | *Oryza sativa Japonica* | F:CAGGAACACCGATAAATGGAA<br>R:TCAATGGTCAGACAAACACCA | 6e - 06 | 52% | Thiolases active site | 0.00 | Polymorphic 400 bp to 870 bp. |
| 3. | gi\|18465444\| | Ribosome biogenesis protein BMS1 -like protein | *Medicago truncatula* | F: ATGAAAGCGAGCCTGATGAT<br>R: TCAAGCAACGATTCTTTCCA | 2E - 26 | 38% | VWFC domain signature | 0.50 | Polymorphic 100 bp to 610 bp. |
| 4. | gi\|18464838\| | DnaJ domain containing protein, expressed | *Oryza sativa* subsp. *japonica* (Rice) | F: CTTTCCGTCAGGAGTTCAGC<br>R: AATGGCAATTCCACGAAGAG | 1E - 36 | 50% | VWFC domain signature | 0.66 | Monomorphic 350 bp to 700 bp. |
| 5. | gi\|18464578\| | Calmodulin-binding protein-like protein | *Arabidopsis thaliana* | F: CAAAGAAAGGCTCCCATCAA<br>R: TTTCTGTGGAGTTGCAGGTG | 2E - 4 | 35% | 4Fe - 4S ferredoxin-type iron - sulfur binding region signature | 0.66 | Polymorphic 620 bp and 750bp. |
| 6. | gi\|18464496\| | B2 protein | *Daucus carota* | F: GGGAAACATGGGAAGAACAA<br>R: CCGGTGTGATTTGCCTTACT | 2E - 69 | 79% | 2Fe - 2S ferredoxin-type iron - sulfur binding region signature | 0.00 | Monomorphic 800 bp to 1100 bp. |
| 7. | gi\|16950303\| | Probable polyprotein allergen Hgg - 14 | *Heterodera glycines* (Soybean cyst nematode worm) | F: GATTCCAATACACGCGCTTT<br>R: CAAAAAGCCCACCAACATTT | 1.5 E - 2 | 80% | 4Fe - 4S ferredoxin-type iron - sulfur binding region signature | 0.49 | Polymorphic 510 bp to 800 bp |
| 8. | gi\|16950250\| | S - adenosylmethionine decarboxylase uORF | *Daucus carota* | F: TCAAGTTAGGGTTCCGTTCG<br>R: GCCGTTTTTCGTATCCTTCA | 5e - 14 | 100% | 2Fe - 2S ferredoxin-type iron-sulfur binding region signature | 0.00 | Monomorphic 500 bp and 600 bp |
| 9. | gi\|16950178\| | Ribosomal RNA processing protein-like protein | *Medicago truncatula* | F: TGTAAATTTGGGCACAATCG<br>R: GGAACATCTTCATCGCCATT | 1e - 23 | 43% | Thiolases active site | 0.54 | Polymorphic 500 bp to 750 bp. |
| 10. | gi\|16949890\| | Kinase-related protein | *Arabidopsis thaliana* | F: GCAGCAAAACCCTAGAGACG<br>R: CTTTCAGGGCACAGAAAAGC | 3e - 11 | 47% | VWFC domain signature | 0.00 | Monomorphic 520 bp to 700 bp. |
| 11. | gi\|16949765\| | Overexpressor of cationic peroxidase 3 | *Arabidopsis thaliana* | F: CAAGGCTTGCTCCGAAATAC<br>R: TCATCTGCAAGTGCTTCCTC | 8.8 | 76% | VWFC domain signature | 0.35 | Polymorphic 680 bp to 900 bp |

**Table 6.** (Continued).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 12. | gi\|16949717\| | NADH dehydrogenase subunit 5 | *Bouvardia glaberrima* | F: GCTCCATCTCCATCATCGTT<br>R: ATTTGGGGTGTGGTTGAGAG | 8.5 | 42% | VWFC domain signature | NA | No amplification |
| 13. | gi\|16949394\| | Resistance to phytophthora 1 protein | *Arabidopsis thaliana* | F: TGCCGACATCCATCTACAAA<br>R: TTGCCGGAGAGCTAGATGTT | 2E - 50 | 88% | Thiolases active site | 0.00 | Monomorphic<br>550 bp to 690 bp. |
| 14. | gi\|16949015\| | Mate efflux family protein | *Arabidopsis lyrata* subsp. *lyrata* | F: GCAAAGTTGCTTCGACATGA<br>R: GCGCTCGACATTCCTAAAAG | 9e - 27 | 87% | 4Fe - 4S ferredoxin-type iron-sulfur binding region signature | 0.00 | Polymorphic<br>780 bp to 1200 bp. |
| 15. | gi\|16948995\| | At3g48070 expression protein | *Arabidopsis thaliana* | F: GACAAATTTCACTGGCAGCA<br>R: TCATCAGCTGCTAACGCATC | 7E - 44 | 47% | Thiolases active site | 0.49 | Polymorphic<br>580 bp to 880 bp. |
| 16. | gi\|16948775\| | Os02g0684400 protein | *Oryza sativa* subsp. *japonica* (Rice) | F: TCTTGCAAGTGTGAGGAGCA<br>R: TTTTGTTGACCAGCGTTGAC | 9E - 19 | 47% | VWFC domain signature | 0.49 | Polymorphic<br>100 bp to 360 bp. |
| 17. | gi\|16948738\| | Squamosa promoter – binding-like protein 3 | *Arabidopsis thaliana* | F: GGATGACTGCACAAGCGATA<br>R: AGCTAAGCGCCTACGACAAC | 4E - 45 | 66% | Thiolases active site | 0.00 | Polymorphic<br>520 bp to 640 bp. |
| 18. | gi\|16948722\| | S - adenosylmethionine decarboxylase uORF | *Daucus carota* | F: TCAAGTTAGGGTTCCGTTCG<br>R: TCCTTCCAAAACCAAATTGC | 4e - 14 | 100% | 2Fe - 2S ferredoxin-type iron-sulfur binding region signature | 0.00 | Monomorphic<br>620 bp to 780 bp. |

**Table 7.** Jaccard's similarity coefficients of *Stevia rebaudiana* based on EST-SSR markers.

| | A | B | St-1 | St-2 | St-3 | St-4 | St-5 | St-6 | NSt-1 | NSt-2 | NSt-3 | CSt-1 | RSt-1 | RSt-2 | RSt-5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **B** | 0.568 | | | | | | | | | | | | | | |
| **St-1** | 0.450 | 0.693 | | | | | | | | | | | | | |
| **St-2** | 0.372 | 0.612 | 0.714 | | | | | | | | | | | | |
| **St-3** | **0.295** | 0.517 | 0.576 | 0.618 | | | | | | | | | | | |
| **St-4** | 0.344 | 0.526 | 0.642 | 0.629 | 0.644 | | | | | | | | | | |
| **St-5** | 0.351 | 0.547 | 0.641 | 0.660 | 0.642 | 0.750 | | | | | | | | | |
| **St-6** | 0.357 | 0.574 | 0.666 | 0.720 | 0.666 | 0.709 | 0.780 | | | | | | | | |
| **NSt-1** | 0.396 | 0.578 | 0.666 | 0.625 | 0.612 | 0.677 | 0.649 | 0.672 | | | | | | | |
| **NSt-2** | 0.360 | 0.559 | 0.672 | 0.690 | 0.619 | 0.629 | 0.627 | 0.736 | 0.733 | | | | | | |
| **NSt-3** | 0.360 | 0.559 | 0.672 | 0.690 | 0.672 | 0.711 | 0.714 | 0.767 | **0.824** | 0.766 | | | | | |
| **CSt-1** | 0.415 | 0.615 | 0.711 | 0.700 | 0.540 | 0.631 | 0.660 | 0.654 | 0.655 | 0.719 | 0.781 | | | | |
| **RSt -1** | 0.389 | 0.542 | 0.573 | 0.586 | 0.578 | 0.612 | 0.583 | 0.580 | 0.634 | 0.615 | 0.666 | 0.616 | | | |
| **RSt -2** | 0.327 | 0.508 | 0.568 | 0.581 | 0.573 | 0.557 | 0.578 | 0.603 | 0.580 | 0.666 | 0.612 | 0.586 | 0.706 | | |
| **Rst -5** | 0.381 | 0.603 | 0.698 | 0.653 | 0.610 | 0.620 | 0.711 | 0.642 | 0.701 | 0.677 | 0.706 | 0.716 | 0.661 | 0.660 | |
| **RSt -6** | 0.395 | 0.551 | 0.588 | 0.571 | 0.508 | 0.666 | 0.538 | 0.596 | 0.543 | 0.578 | 0.551 | 0.576 | 0.618 | 0.448 | 0.596 |

**Table 8.** Dissimilarity coefficients of EST-SSR analysis generated by DARwin5 ver. 5.0.155.

| | A | B | St-1 | St-2 | St-3 | St-4 | St-5 | St-6 | NSt-1 | NSt-2 | NSt-3 | CSt-1 | RSt-1 | RSt-2 | RSt- 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **B** | 0.211 | | | | | | | | | | | | | | |
| **St-1** | 0.311 | 0.166 | | | | | | | | | | | | | |
| **St-2** | 0.355 | 0.211 | 0.155 | | | | | | | | | | | | |
| **St-3** | **0.477** | 0.311 | 0.277 | 0.233 | | | | | | | | | | | |
| **St-4** | 0.422 | 0.300 | 0.222 | 0.222 | 0.233 | | | | | | | | | | |
| **St-5** | 0.388 | 0.266 | 0.211 | 0.188 | 0.222 | 0.144 | | | | | | | | | |
| **St-6** | 0.400 | 0.255 | 0.200 | 0.155 | 0.211 | 0.177 | 0.122 | | | | | | | | |
| **NSt-1** | 0.388 | 0.266 | 0.211 | 0.233 | 0.266 | 0.211 | 0.222 | 0.211 | | | | | | | |
| **NSt-2** | 0.433 | 0.288 | 0.211 | 0.188 | 0.266 | 0.255 | 0.244 | 0.166 | 0.177 | | | | | | |
| **NSt-3** | 0.433 | 0.288 | 0.211 | 0.188 | 0.222 | 0.188 | 0.177 | 0.144 | **0.111** | 0.155 | | | | | |
| **CSt-1** | 0.344 | 0.222 | 0.166 | 0.166 | 0.311 | 0.233 | 0.200 | 0.211 | 0.222 | 0.177 | 0.133 | | | | |
| **RSt-1** | 0.400 | 0.300 | 0.288 | 0.266 | 0.300 | 0.266 | 0.277 | 0.288 | 0.255 | 0.277 | 0.233 | 0.255 | | | |
| **RSt-2** | 0.433 | 0.311 | 0.277 | 0.255 | 0.288 | 0.300 | 0.266 | 0.255 | 0.288 | 0.222 | 0.266 | 0.266 | 0.188 | | |
| **Rst-5** | 0.377 | 0.233 | 0.177 | 0.200 | 0.255 | 0.244 | 0.166 | 0.222 | 0.188 | 0.211 | 0.188 | 0.166 | 0.222 | 0.211 | |
| **RSt-6** | 0.322 | 0.244 | 0.233 | 0.233 | 0.311 | 0.188 | 0.266 | 0.233 | 0.288 | 0.266 | 0.288 | 0.244 | 0.233 | 0.355 | 0.233 |

signature, EGF-like domain signature 1, and 2Fe-2S ferredoxin-type iron-sulfur binding region signature.

The detection of a higher level of polymorphism, i.e. 61.11%, using SSRs was in agreement with reports by some other researchers: Pinto et al. (2004) recorded 68.6% polymorphism in sugarcane and Chen et al. (2006) reported 59.3% polymorphism in citrus. The high PIC presented would facilitate QTL identification and marker-assisted selection due to the association with functional regions of the genome.

The present study clearly revealed a great level of variation among all the collections. Dendrogram clustering in the present study based on UPGMA

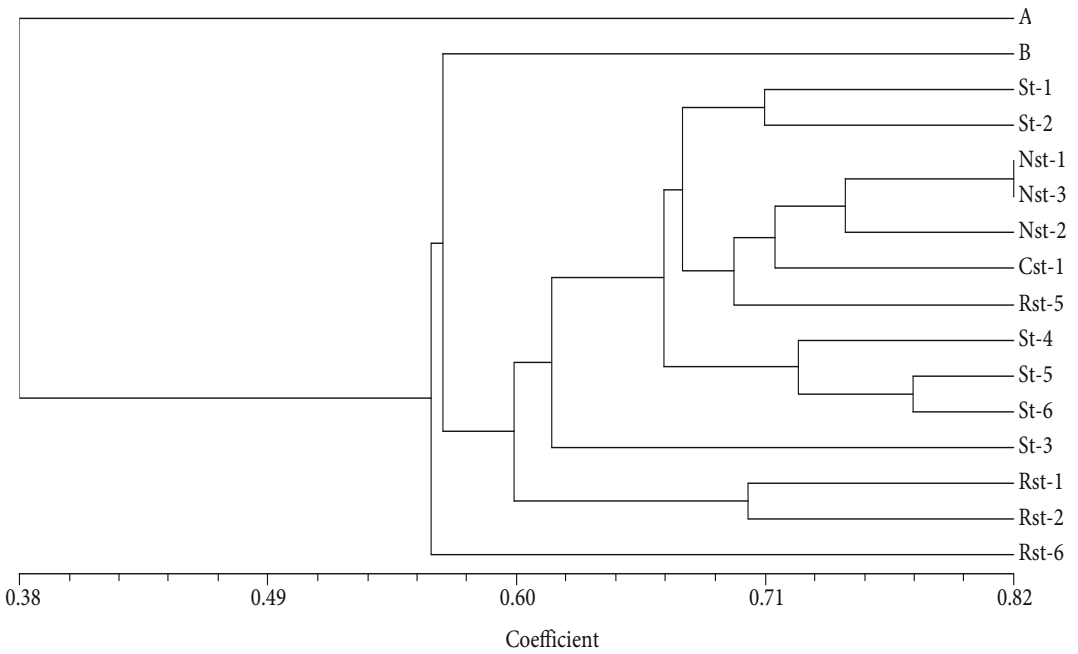


**Figure.** Dendrogram derived using EST-SSR primers.

grouped various stevia collections based on their response to EST-SSR primers that until now had done based only on their morphology.

The present study is an initiative in the direction of development of EST-based microsatellite markers that can be used for genetic studies of *Stevia rebaudiana*. It is hoped that with the increasing emphasis on computational biology large EST resources will become available for stevia in the near future for designing various breeding programs.

## References

Adams MD, Kelley JM, Gocayne JD (1992). Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651–1656.

Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V and 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.

Ahmed MS Dobberstein RH (1982) *Stevia rebaudiana* II. High performance liquid chromatographic separation and quantitation of stevioside, rebaudioside-A and rebaudioside-C. J Chromatogr A 236: 519–522.

Alwala S, Suman A, Arro JA, Veremis JC (2006). Target region amplification polymorphism (TRAP) for assessing genetic diversity in sugarcane germplasm collections. Crop Sci 46: 448–455.

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

Brandle JE, Richman A, Swanson AK, Chapman BP (2002). Leaf ESTs from *Stevia rebaudiana*: a resource for gene discovery in diterpene synthesis. Plant Mol Biol 50: 613–622.

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

Cloutier S, Niu ZX, Datla R, Duguid S. 2009. Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). Theor Appl Genet 119: 53–63.

Dacome AS, Da Silva CC, Da Costa CEM, Fontana JD, Adelmann J, Da Costa SC (2005). Sweet diterpene glycosides balance of a new cultivar of *Stevia rebaudiana* (Bert.) Bertoni: isolation and quantitative distribution by chromatographic, spectroscopic, and electrophoretic methods. Process Biochem 40: 3587–3594.

Durand J, Bodenes C, Chancerel E, Frigerio JM (2010). A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. BMC Genomics 11: 570.

Ellis JR, Burke JM (2007). EST-SSRs as a resource for population genetic analyses. Hered 99: 125–132.

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM (1999). Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 9: 950–959.

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 Genome 270: 315–323.

Gupta PK, Varshney RK (2000). The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphyt 113: 163–185.

Heikal AH, Badawy OM, Afaf HM (2008). Genetic relationships among some Stevia (*Stevia rebaudiana* Bertoni) accessions based on ISSR analysis. Res J Cell Mol Biol 2: 1–5.

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

Igarashi M, Abe Y, Hatsuyama Y, Ueda T, Fukasawa-Akada T, Kon T, Kudo T, Sato T, Suzuki M (2008). Linkage maps of the apple (*Malus* × *domestica* Borkh.) cvs 'Ralls Janet' and 'Delicious' include newly development EST markers. Mol Breed 22: 95–118.

Ishima N, Katayama O (1976). Sensory evaluation of stevioside as a sweetener. Report of the Nat Food Res Inst 31: 80–85.

Jiao Y, Jia HM, Li XW, Chai ML, Jia HJ, Chen Z, Wang GY, Chai CY, Weg EV, Gao ZS (2012). Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). BMC Genomics 13: 201–216.

Kennelly EJ (2002). Sweet and non-sweet constituents of *Stevia rebaudiana* (Bertoni). Bertoni. In: Kinghorn AD, editor. Stevia, the genus Stevia: medicinal and aromatic plants - industrial profiles. London, UK: Taylor and Francis, pp. 68–85.

Kinghorn AD, Soejarto DD (1985). Current status of stevioside as a sweetening agent for human use. In: Wagner H, Hikino H, Farnsworth NR editors. Economic and medicinal plant research. New York, NY, USA: Academic Press, pp. 1–52.

Kumar YH, Ranjan A, Asif MH, Mantri S (2011). EST-derived SSR markers in *Jatropha curcas* L. development, characterization, polymorphism, and transferability across the species/genera. Tree Genet Genomes 7: 207–219.

Lima LS, Gramacho KP, Pires JL, Clement D (2010). Development, characterization, validation, and mapping of SSRs derived from *Theobroma cacao* L. - *Moniliophthora perniciosa* interaction ESTs. Tree Genet Genomes 6: 663–676.

Liu ZH, Anderson JA, Hu J, Friesen TL (2005). A wheat intervarietal linkage map based on microsatellite and target region amplified polymorphism markers and its utility for detecting quantitative trait loci. Theor Appl Genet 111: 782–794.

Masoudi-Nejad A, Koichiro T, Shuichi K, Yuki M, Masanori S, Masumi I, Minoru K, Takashi E, Susumu G (2006). EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res 34: 459–462.

Miklas PN, Hu J, Grunwald NJ, Larsen KM (2006). Potential application of TRAP (targeted region amplified polymorphism) markers for mapping and tagging disease resistance traits in common bean. Crop Sci 46: 910–916.

Morand ME, Brachet S, Rossignol P, Dufour J, Frascaria-Lacoste N (2002). A generalized heterozygote deficiency assessed with microsatellites in French common ash populations. Mol Ecol 11: 377–385.

Pashley CH, Ellis JR, McCauley DE, Burke JM (2006). EST databases as a source for molecular markers: lessons from *Helianthus*. J Hered 97: 381–388.

Pinto LR, Oliveira KM, Ulian EC, Garcia AAF, de Souza AP (2004). Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. Genome 47: 795–804.

Rallo P, Dorado G, Martin A (2000). Development of simple sequence repeats (SSRs) in olive tree (*Olea europaea* L.). Theor Appl Genet 101: 984–989.

Richman A, Swanson A, Humphrey T, Chapman R, McGarvey B, Pocs R, Brandle JE (2005). Functional genomics uncovers three glucosyltransferases involved in the synthesis of the major sweet glucosides of *Stevia rebaudiana*. Plant J 41: 56–67.

Rohlf FJ (2000). NTSYS - pc numerical taxonomy and multivariate analysis version 2.0. Applied New York: Biostatics Inc., p. 25.

Ronning C, Stegalkina S, Ascenzi R (2003). Comparative analyses of potato expressed sequence tag libraries. Plant Physiol 131: 419–429.

Scott KD, Eggler P, Seaton G, Rosseto M, Ablett EM, Lee LS, Henry RJ (2000). Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100: 723–726.

Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Lowe AJ, Gibby M (2003). How much effort is required to isolate nuclear microsatellites from plants? Mol Ecol 12: 1339–1348.

Starratt AN, Kirby CW, Pocs R, Brandle JE (2002). Rebaudioside-E, a diterpene glycoside from *Stevia rebaudiana*. Phytochem 59: 367–370.

Temnykh S, Declerk G, Lukashova A, Lipovich L, Cartinhour S, Mccouch SR (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.

Vaidya E, Kaur R, Bhardwaj SV (2012). Datamining of ESTs to develop dbEST-SSR for use in polymorphism study of cauliflower (*Brassica oleracea* var. *botrytis*). J Hortic Sci Biotechnol 87: 57–63.

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 Literature 7: 537–546.

Verma M, Arya L (2008). Development of EST-SSRs in watermelon (*Citrullus lanatus* var. *lanatus*) and their transferability to *Cucumis* spp. J Hortic Sci Biotechnol 83: 732–736.

Vosman B, Arens P (1997). Molecular characterization of GATA/GACA microsatellite repeats in tomato. Genome 40: 25–33.

Wijarat P, Keeratinijakal V, Toojinda T, Vanavichit A, Tragoonrung S (2012). Genetic evaluation of *Andrographis paniculata* (Burm. f.) Nees revealed by SSR, AFLP and RAPD markers. J Med Plants Res 6: 2777–2788.

Woodhead M, McCallum S, Smith K, Cardle L, Mazzitelli L, Graham J (2008). Identification, characterisation and mapping of simple sequence repeat (SSR) markers from raspberry root and bud ESTs. Mol Breed 22: 555–563.

Yao Y, Ban M, Brandle J (1999). A genetic linkage map for *Stevia rebaudiana*. Genome 42: 657–661.

Yi G, Lee JM, Lee S, Choi D, Kim BD (2006). Exploitation of pepper EST-SSRs and an SSR-based linkage map. Theor Appl Genet 114: 113–130.

Zane L, Bargelloni L, Patarnello T (2002). Strategies for microsatellite isolation: a review. Mol Ecol 11: 1–16.

Zeid M, Schon C, Link W (2003). Genetic diversity in recent elite faba bean lines using AFLP markers. Theor Appl Genet 107: 1304–1314.

Zeng S, Xiao G, Guo J, Fei Z, Xu Y, Roe BA, Wang Y (2010). Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. BMC Genomics 11: 94.

Zhu Y, Hao Y, Wang K, Wu C, Wang W, Qi J, Zhou J (2009). Analysis of SSRs information and development of SSR markers from walnut ESTs. J Fruit Sci 26: 394–398.