Large-scale advances in SSR markers with high-throughput sequencing in Euphorbia fischeriana Steud

Hui Li, Yanshi Ma, Fangyi Pei, Haiyan Zhang, Jicheng Liu, Ming Jiang *

Department of Science and Technology, Qiqihar Medical University, Qiqihar 161006, Heilongjiang, China

ARTICLE INFO

Article history:
Received 22 April 2020
Accepted 24 November 2020
Available online 3 December 2020

Keywords:
Cross-species
EST-SSRs
Euphorbia fischeriana Steud
Genetic diversity analysis
High-throughput sequencing
Medicinal herb
Molecular marker-assisted breeding
Molecular markers
SSR

ABSTRACT

Background: Euphorbia fischeriana Steud is a very important medicinal herb and has significant medical value for healing cancer, edema and tuberculosis in China. The lack of molecular markers for Euphorbia fischeriana Steud is a dominant barrier to genetic research. For the purpose of developing many simple sequence repeat (SSR) molecular markers, we completed transcriptome analysis with the Illumina HiSeq 2000 platform.

Results: Approximately 9.1 million clean reads were acquired and then assembled into approximately 186.3 thousand nonredundant unigenes, 53,146 of which were SSR-containing unigenes. A total of 76,193 SSR loci were identified. Of these SSR loci, 28,491 were detected at the terminal position of ESTs, which made it difficult to design SSR primers for these SSR-containing sequences, and the residual SSRs were thus used to design primer pairs. Analyzing the results of these markers revealed that the mononucleotide motif A/T (44,067, 57.83% of all SSRs) was the most abundant, followed by the dinucleotide type AG/CT (9430, 12.38%). Using 100 randomly selected primer pairs, 77 primers were successfully amplified in Euphorbia fischeriana Steud, and 79 were successfully amplified in three other related species. The markers developed displayed relatively high quality and cross-species transferability.

Conclusions: The large number of EST-SSRs exploited successfully in Euphorbia fischeriana Steud for the first time could provide genetic information for research on linkage maps, variety identification, genetic diversity analysis, and molecular marker-assisted breeding.


© 2020 Pontificia Universidad Católica de Valparaíso. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Euphorbia fischeriana Steud is a traditional Chinese medicinal plant belonging to the Euphorbiaceae family (spurge family) and has a close relationship with other Euphorbia species, including Euphorbia ehrnacteolate Hayata and Euphorbia esula Linn, as well as Stellera chamaejasme Linn. The medicinal part of the plant is called “Lang-Du” and is mainly used in Northeast China as a traditional Chinese medicinal plant to treat cancer, edema and tuberculosis [1]. However, wild Euphorbia fischeriana Steud resources are extremely disturbed, and the germplasm population has decreased sharply because of overexploitation [2].

Recently, a number of counterfeit plants have been used as Euphorbia fischeriana Steud to treat associated diseases. Different species with the same name are found in Thymelaeaceae, Euphorbiaceae, Araceae, Ranunculaceae and other families [3,4]. However, research indicates that species and other factors significantly influence the chemical components and medicinal effect of plants [5]. This provides a chance to characterize the genetic basis of key traits in Euphorbia fischeriana Steud and enhance the quality of this plant through molecular biotechnology [6]. The genetic diversity of Euphorbia fischeriana Steud and other medicinal plants has typically been assessed by random amplified polymorphic DNA (RAPD) [7] and amplified fragment length polymorphism (AFLP) [8] technologies.

The genetic diversity of Euphorbia fischeriana Steud populations in different habitats was previously researched with inter-simple sequence repeat (ISSR) markers [9]. However, these molecular markers have several limitations, such as poor reproducibility and incomplete dominance. Instead, simple sequence repeat (SSR) molecular markers display many benefits due to their rich content distribution, abundant polymorphism, reliable neutrality, high repeatability and codominance [10]. Therefore, SSR markers...
have been widely adopted to research genetic diversity, geographical variation, molecular marker-assisted breeding, and variety identification in many vegetative plants [11,12,13,14]. SSRs are normally obtained by genomic sequencing of SSR-containing sections or library construction of expressed sequence tags (ESTs). However, these methods are considered laborious and time-consuming, and genomic SSR markers mainly originate from intergenic DNA fragments, which have unclear locations in the genome. In contrast, transcriptome analysis with high-throughput DNA sequencing technologies on a large scale has been used to rapidly identify genic SSRs with high efficiency. Thus, many SSR molecular markers have been developed and analyzed with this technology in many herbal plants [15,16,17,18,19]. However, few SSR markers for *Euphorbia* have been published to date. Although studies have characterized the transcriptome of *Euphorbia fischeriana* Steud, SSR loci in this species have not been reported [20]. The purpose of the present research was to exploit the EST-SSR molecular markers in a Chinese herbal plant. A *Euphorbia fischeriana* Steud cultivar called “Lang-Du-Da-Ji” was chosen for SSR locus identification and marker development from unigenes assembled on the basis of transcriptome analysis. Moreover, the cross-species transferability of *Euphorbia fischeriana* and Steud SSR markers was examined.

## 2. Materials and methods

### 2.1. Plant materials and RNA and DNA extraction

A cultivar of *Euphorbia fischeriana* Steud, a very important Chinese medicinal herb, was used for transcriptome analysis. Three fresh roots of this plant were collected 150 d after planting, frozen in liquid nitrogen and preserved at –80°C immediately for further study. Total RNA was extracted with TRIzol reagent (TransGen Biotech, Beijing, China) on the basis of the product description and used for high-throughput transcriptome sequencing. Fresh leaves of *Euphorbia fischeriana* Steud and three other species (*Stellera chamaejasme* Linn, *Euphorbia ebracteolate* Hayata and *Euphorbia esula* Linn) were gathered, and DNA samples were isolated using a Plant Genomic DNA Kit (Tiangen, China) based on the user instructions.

### 2.2. cDNA library construction, sequencing, and assembly

Illumina analysis was performed by Novogene Bioinformatics Technology Co., Ltd., Beijing, China (www.novogene.cn). The cDNA library of *E. fischeriana* was constructed with suitable RNA fragments (approximately 180 bp). After that, the paired-end strategy was carried out with the Illumina HiSeq™ 2000 platform. To obtain high-quality reads, we removed the empty reads, N-containing reads and common adaptor sequences from the raw sequences. After that, *de novo* assembly was performed with these filtered clean reads using Trinity [21].

### 2.3. Gene annotation of SSR-containing unigenes

For the purpose of annotating the assembled SSR-containing unigenes, which were first probed from different databases, including the public NCBI nonredundant protein (Nr) database, the Swiss-Prot protein sequence database, the NCBI nonredundant nucleotide sequence (Nt) database, the Gene Ontology (GO) database, the Kyoto Encyclopedia of Genes and Genomes (KEGG) knowledge database, the Pfam database and the Cluster of Orthologous Groups (COG) database. GO functional classification was executed with the software WEGO [22,23]. GO terms were separated into three components: biological processes (BPs), cellular components (CCs) and molecular functions (MFs).

### 2.4. Discovery of SSR loci and development of markers

The SSR loci were identified from the managed unigenes with Microsatellite Identification Tool (MISA) software (http://pgrc.ipk-gatersleben.de/misa/). The default criteria were utilized to select a minimum of ten repeats for mononucleotide motifs, six for dinucleotide motifs and five for the other four motif types (trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide). SSR primers flanking the putative SSR loci were designed based on the EST sequences with Primer 3.0. The standard parameters related to target amplified fragment lengths, annealing temperatures, GC contents and primer sizes were set according to Mishra et al. [24].

### 2.5. Amplification of SSR markers

*Euphorbia fischeriana* Steud and 3 other species (*Euphorbia ebracteolate* Hayata, *Stellera chamaejasme* Linn and *Euphorbia esula* Linn) were selected to assess one hundred chosen SSR primers (LDDL0001-LDDL0100) and their cross-species universality. The PCR amplification system included a total volume of 20 μl consisting of 2 μl 10 × PCR buffer, 0.8 μl (10 mmol/L) dNTPs, 0.2 μl (5 U/μl) DNA polymerase (TransGen Biotech, Beijing, China), 1.0 μl (7.5 mmol/L) each SSR primer, 3 μl genomic DNA (50 ng/μl), and 12 μl deionized distilled water. The PCR program was as follows: 5 min at 95°C, followed by 35 cycles of 30 s at 95°C, 1 min at 59°C, 1 min at 72°C, and a sufficient extension step at 72°C for 10 min and a final cooling step of 5 min at 4°C. The PCR analysis and silver staining method were as described by Meng et al. [25].

## 3. Results

### 3.1. High-throughput sequencing and read assembly

The transcriptome data of *Euphorbia fischeriana* Steud were analyzed by the Illumina HiSeq 2000 sequencing platform, and 9.1 million raw reads were acquired from the library and used for assembly. The dataset included 186,384 nonredundant unigenes with a total length of 270.6 Mb and an average length of 1452 bp (Table 1). Among these unigenes, there were 8861 (4.8%) with a length of not more than 300 bp, 26,065 (14.0%) with a length between 301 and 500 bp, 52,210 (28.0%) with a length between 501 and 1000 bp, 54,429 (29.2%) with a length between 1001 and 2000, and 44,819 (24.0%) with a length greater than 2000 bp (Fig. 1).

### 3.2. Gene annotation and classification of SSR-containing unigenes

The SSR-containing unigene sequences were validated and annotated with seven different public databases. Among these 53,146 unigenes, 42,395 (79.77%) were related to known protein information in the NR database; 37,898 (71.31%), in the NT database; 35,136 (65.96%), in the Swiss-Prot database; 22,645 (43.57%), in the Pfam database; 12,805 (24.13%), in the COG database; 10,320 (19.30%), in the KEGG database; and 9,688 (18.11%), in the GO database. In total, 31,147 unigenes were annotated with these seven public databases. Among these annotated unigenes, 21,380 (69.03%), 14,841 (47.46%), 13,441 (43.25%), 10,639 (34.00%), 7,237 (23.02%), 4,947 (16.19%), and 5,437 (17.29%) were annotated with the NR, NT, Swiss-Prot, Pfam, COG, KEGG, and GO databases, respectively.

### Table 1

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear reads</td>
<td>91,413,326</td>
</tr>
<tr>
<td>Clean bases (Gb)</td>
<td>13.72</td>
</tr>
<tr>
<td>Unigene number</td>
<td>186,384</td>
</tr>
<tr>
<td>Total length of unigene (Mb)</td>
<td>270.6</td>
</tr>
<tr>
<td>Average length of unigene (bp)</td>
<td>1452</td>
</tr>
<tr>
<td>Unigene number annotated</td>
<td>144,356 (77.43%)</td>
</tr>
</tbody>
</table>
3.3. Frequency and distribution of SSRs

All 186,384 clustered unigenes were used to check possible SSR loci in this study, and 53,146 unigenes including 76,193 SSR loci were checked. The frequency of detection for *Euphorbia fischeriana* Steud was 1 SSR per 3.55 kb of EST sequence. Among the unigenes, 6224 SSRs presented a compound formation, and 16,101 included more than one SSR. The mononucleotide motif was the most abundant type (44,810, 58.81%) among the illustrated SSRs, followed by the di- (14,245, 18.69%), tri- (14,226, 18.67%), hexa- (1651, 2.17%), tetra- (696, 0.91%), and pentanucleotide (565, 0.74%) types (Table 2).

In all SSR loci, 185 different motif types were recognized. The most abundant mononucleotide type was A/T (44,067, 57.83% of all SSRs), and there were 743 C/G nucleotides in total. AG/CT was the most abundant dinucleotide type (9430, 12.38%), followed by AT/AT (3373, 4.43%) and AC/GT (1417, 1.86%). Of the trinucleotides, AAAG/CTTT was the most common type (6404, 8.41%), followed by AAC/ATG (1897, 2.49%) and ACC/GGT (1604, 2.11%). The AAAG/CTTT (384, 0.50%) motif was the most frequent tetranucleotide type, and the most common pentanucleotides and hexanucleotides were AAAAG/CTTTT (96, 0.13%) and AAAAGG/CTTTTT (29, 0.04%), respectively (Fig. 3). Most of the SSR repeat numbers ranged from 5-20. Of the dinucleotide motif types, 6 repeats was the most common (8116 SSRs), followed by 7 repeats. Of the trinucleotides, the SSR numbers decreased with decreasing repeat numbers. Most of the SSR lengths (47.2%) ranged from 12 to 18 bp, followed by 19 to 25 bp (13.9%), and 2455 (3.2%) were longer than 30 bp.

3.4. SSR marker development

Of the 76,193 SSR loci, 28,491 were positioned at the ends of an EST, which made it difficult to design primer pairs related to regions surrounding these loci. We designed 1000 primer pairs for analyzing markers randomly selected from these sequences. These SSR markers were assigned individual numbers (HQ0001-HQ1000).

3.5. Quality estimation of the SSR markers

We amplified 100 SSRs to estimate the applicability of these developed SSR markers with PCR technology. The result was considered successful if a single fragment was amplified with PCR and/or one of the fragments displayed an accurate length when there were several fragments amplified by a primer pair.

In the end, 77 primer pairs achieved exact amplification. Among these primer pairs, 33 produced multiple fragments, and 44 resulted in a single fragment from PCR amplification. Among the sole-amplicon primers, 21 fragments showed accurate sizes, 21 showed longer lengths than expected, and the other 2 showed smaller fragments than expected.

3.6. Transferability evaluation of SSR markers

The species universality of these 100 SSR markers was estimated in three different species. For these 100 markers, 45, 72 and 53 primer pairs were used to obtain PCR fragments from the three species (*Stellera chamaejasme* Linn, *Euphorbia ebracteolate* Hayata and *Euphorbia esula* Linn), respectively (Fig. S1). PCR amplicons obtained by 12 primer pairs from *Euphorbia ebracteolate* Hayata, 28 primer pairs from *Stellera chamaejasme* Linn and 15 primer pairs from *Euphorbia esula* Linn contained multiflampsions. For these primer pairs, it was complicated to distinguish specific fragments from nonspecific fragments because the accurate lengths were unknown. Thus, the SSRs were considered successful and transferable because of special sites combined with the genomic resources of the other species. Consequently, 44 markers showed high universality in all three other species, and 79 could amplify DNA products in at least one of those species.

4. Discussion

In medicinal plants, expressed sequence tags (ESTs) are regarded as an effective strategy for aspects of gene discovery, genomic sequence analysis and important genomic functional annotation, including genetic map construction, molecular marker-assisted breeding, variety identification, microarrays and gene expression profiling. Hence, many researchers have constructed and sequenced cDNA libraries for model plants, such as *Prosopis cineraria* [26], *manihot* [27] and maize [28], to obtain an EST database. Traditional sequencing strategies are not only time-consuming but also costly, which has greatly limited the number of ESTs developed in nonmodel plants, especially Chinese medicinal herbs. With the application of less time-consuming and less expensive deep-sequencing methods, many ESTs for nonmodel medicinal herbs have become available. To date, the sequencing of several nonmodel plants has been finished, and hundreds of thousands of ESTs have been obtained [29,30,31].

The lack of SSR molecular markers is an important hurdle to genomic analysis of *Euphorbia* species, with only a few SSR markers having been reported, making genetic map construction and species identification difficult. In this research, 1000 SSR primer pairs were used to obtain ESTs based on transcriptome analysis. One hundred SSRs were amplified to evaluate the quality of these primers, and 77 primer pairs finally amplified the target fragments (77%). Because these SSR loci originated from EST sequences, 23% of the SSRs were unsuccessful, mainly because the primers were positioned in introns and/or at splice sites within target fragments.
For the 100 SSRs, 44 primer pairs amplified only one amplicon, indicating that 44% of the SSRs obtained in the research showed high usefulness. Thirty-three primer pairs resulted in nonspecific fragments, which may be a result of the highly repetitive nature of herbal genome sequences. Even so, the number of nonspecific fragments resulting from the SSR primers was also far smaller than the numbers resulting from AFLP and RAPD techniques. According to the location information, we obtained 10 SSR markers linked to key genes related to biosynthetic pathway of prostratin (3 linked to 3-hydroxy-3-methylglutaryl-CoA reductase, 2 to 1-deoxy-D-xylulose 5-phosphate reductoisomeras, 1 to 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, 1 to 1-deoxy-D-xylulose 5-phosphate synthase, 1 to 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase, 1 to farnesyl pyrophosphate synthase, 1 to Geranylgeranyl pyrophosphate synthase), which is one particular tigliane diterpene in this plant [20]. Thus, the markers acquired in this study might provide an important resource in future studies for molecular marker-assisted selection. The 3.55-kb interval discovered in Euphorbia EST-SSRs in this study showed that the frequency of EST-SSRs was equivalent to the frequency found in tea (3.5 kb) [32] and higher than that in most other plants [33,34]. Among the repeat motifs, mononucleotide repeats were the most abundant (58.81%), followed by dinucleotides (18.69%) and trinucleotides (18.67%). The number of dinucleotides was the same as that of trinucleotides but not that of mononucleotide motifs, which is contradictory to previous
results in other plants [35,36]. The different results related to motif frequency, SSR number and average distance could be explained by differences in search criteria and the abundance of ESTs in the databases obtained in previous studies [37]. The A/T, AG/CT and AAG/CTT types were the most abundant, accounting for 57.83%, 12.38% and 8.41 of all the repeat types, respectively. Within GC-rich motifs, CG/CG and CCG/CGG were detected, accounting for 0.03% and 0.65%, respectively. The same result, i.e., a low abundance of GC repeats, was obtained in dicot plants [14].

Among these 100 SSRs from *Euphorbia fischeriana* Steud, 79 displayed universality in at least one of three other species, 44 of which showed universality in all three species. High transferability has also been found in many other plants, including *Chrysanthemum nankinense* [38], *Dysosma versipellis* [39], and *Pinus densiflora* [40].

![Fig. 3. Frequency characterize of EST-SSRs on the basis of motif types.](image)

Table 2
Description of SSR in transcriptome of *Euphorbia fischeriana* Steud.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total unigene number</td>
<td>186,384</td>
</tr>
<tr>
<td>SSR loci</td>
<td>76,193</td>
</tr>
<tr>
<td>SSR-contained unigenes</td>
<td>53,146</td>
</tr>
<tr>
<td>Unigenes contained more than one SSR</td>
<td>16,101</td>
</tr>
<tr>
<td>Number of SSRs presented a compound formation</td>
<td>6224</td>
</tr>
<tr>
<td>SSR motif of mononucleotide</td>
<td>44,810</td>
</tr>
<tr>
<td>SSR motif of dinucleotide</td>
<td>14,245</td>
</tr>
<tr>
<td>SSR motif of trinucleotide</td>
<td>14,226</td>
</tr>
<tr>
<td>SSR motif of tetrinucleotide</td>
<td>1651</td>
</tr>
<tr>
<td>SSR motif of pentanucleotide</td>
<td>696</td>
</tr>
<tr>
<td>SSR motif of hexanucleotide</td>
<td>565</td>
</tr>
</tbody>
</table>

**Financial support**
This work was supported by the Fundamental Research Funds for Education Department of Heilongjiang Province (no. 2017-KYYWF-0107).

**References**


