Tropical Journal of Pharmaceutical Research January 2019; 18 (1): 75-84
**ISSN:** 1596-5996 (print); 1596-9827 (electronic)


Available online at http://www.tjpr.org
**http://dx.doi.org/10.4314/tjpr.v18i1.12**

**Original Research Article**

# Comparative transcriptome analysis and simple sequence repeat marker development for two closely related Isodon species used as 'Xihuangcao' herbs

**Shanshan Huang[1,2], Weiming Hu[1], Shaohua Zeng[1], Xiaolu Mo[2], Ying Wang[1]***
*[1]Guangdong Provincial Key Laboratory of South China Agricultural Plant Molecular Analysis and Genetic Improvement, and Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510610, [2]Guangdong Food and Drug Vocational College, Guangzhou 510520, PR China*

**For correspondence: **Email:** yingwang@scib.ac.cn; **Tel:** +86-13797038842*

*Sent for review: 3 September 2018* *Revised accepted: 14 December 2018*

## Abstract

***Purpose:*** *To facilitate the molecular identification of original plants, resolve taxonomic problems and identify standards for 'Xihuangcao'-based products on the market.*
***Methods:*** *A transcriptomic analysis of two closely related species, i.e., Isodon serra (Maxim.) (IS) and I. lophanthoides (Buch.-Ham. ex D. Don) Hara, was conducted by using the Illumina HiSeq 2500 platform, and expressed sequence tag-derived simple sequence repeat (EST-SSR) markers were developed based on these transcriptomes.*
***Results:*** *In total, 149,650 and 103,221 contigs were obtained, with N50 values of 1,400 and 1,516, from the IS and I. lophanthoides RNA-Seq datasets, respectively. These contigs were clustered into 107,777 and 68,220 unigenes, which were functionally annotated to identify the genes involved in therapeutic components. In total, 14,138 and 11,756 EST-SSR motifs were identified, and of these motifs, 7,453 and 6,428 were used to design primers for IS and I. lophanthoides, respectively. After PCR verification and fluorescence-based genotyping, 24 SSR markers with bright bands, high polymorphism, and single amplification were obtained and used to identify closely related Isodon species/varieties.*
***Conclusion:*** *These data could help herbal scientists identify high-quality herbal plants and provide a reference for genetic improvement and population genetic and phylogenetic studies investigating 'Xihuangcao' herbs.*

***Keywords:*** *Xihuangcao, Transcriptome, EST-SSR, Molecular markers*



**Tropical Journal of Pharmaceutical Research is indexed by Science Citation Index (SciSearch), Scopus, International Pharmaceutical Abstract, Chemical Abstracts, Embase, Index Copernicus, EBSCO, African Index Medicus, JournalSeek, Journal Citation Reports/Science Edition, Directory of Open Access Journals (DOAJ), African Journal Online, Bioline International, Open-J-Gate and Pharmacy Abstracts**

## INTRODUCTION

*Xihuangcao* (named after its yellow leaf juice) is a traditional Chinese medicinal plant that is well known for its antibacterial, anti-inflammatory and antitumor activities [1]. This herb is mainly used to treat acute icteric hepatitis, cholecystitis, hepatitis, enteritis, sphagitis, and gynecopathy [2]. '*Xihuangcao'* on the herbal medicinal market is mainly derived from the following two species and three varieties: *Isodon serra* (Maxim.) Kudo (IS, the research material in this paper), *I. lophanthoides* (Buch.-Ham. ex D. Don) Hara, *I. lophanthoides var. lophanthoides* (Buch.-Ham.

ex D. Don) H. Hara (ILL, the research material in this paper), *I. lophanthoides var. gerardiana* (Benth.) Hara and *I. lophanthoides var. graciliflora* (Benth.) Hara [3]. These species/varieties are similar in morphology but have slight differences [4].

To date, nearly 300 compounds, including mainly volatile oils, diterpenes, triterpenes, polyphenols, flavonoids, ceramide compounds and active polysaccharides, have been isolated from *Xihuangcao* herbs and identified [1,5]. However, the concentrations of these compounds differ among various varieties of Xihuangcao [6-7]. For example, most diterpenes in IS belong to the abeo-abietanoid type (*lophanthoides* A-F), which has potent antibacterial and antitumor activities [8-9], while the diterpenes in ILL mainly belong to the *ent*-kaurene type, which has high antioxidant and antiviral activities [10]. Thus, the most important part of *Xihuangcao* herb exploitation is the correct identification of the original plant.

Thus far, considerable efforts have been exerted to identify 'Xihuangcao' herbs by morphology, microstructure, chemical composition (chemical markers) and molecular polymorphism [11-13]. However, the inadequacy of markers has led to considerable taxonomic confusion; in particular, only a few molecular markers (e.g., random amplified polymorphic DNA [RAPD] and genomic simple sequence repeat (SSRs) have been developed and used for *Isodon* taxonomy [13,14]. Furthermore, the lack of genomic resources has severely hampered the development of population genetic, morphological, phylogenetic and drug studies. Therefore, in this paper, a transcriptomic analysis of IS and ILL was conducted by using the Illumina platform. The aim of this study was to obtain EST data from these two species and discover and validate SSR markers to authenticate *Xihuangcao* herb-related species/variants.

## EXPERIMENTAL

### Plant materials

The plant materials used in this study are shown in Table 1. All materials were obtained from the Guangdong Research Institute of Traditional Chinese Medicine (Guangzhou, Guangdong 51052, China) in October 2014. The herbs were botanically authenticated by Prof. Huagu Ye at South China Botanical Garden, Chinese Academy of Sciences (Guangzhou, Guangdong 510650, China). A voucher specimen (No. 14058) was deposited at the Herbarium of South China Botanical Garden.

**Table 1:** Seven 'Xihuangcao' herb samples used in this study

| Sample ID | Genus | Species/variety |
|---|---|---|
| A | *Isodon* | *I. lophanthoides var. lophanthoides* |
| B | | *I. serra* |
| C | | *I. lophanthoides* var. *graciliflora* Hara |
| D | | *I. lophanthoides* var. *gerardiana* (Benth.) Hara |
| E | | *I. nervosus* (Hemsley) Kudô, I. walkeri (Arn.) H. Hara |
| F | | *I. walker* (Arn.) H. Hara |
| G | | *I. coetsa* (Buchanan-Hamilton ex d. Don) Kudô |

### Transcriptome sequencing and assembly

The fresh leaves, roots, stems and whole flowers from several individual plants were frozen in liquid nitrogen for immediate RNA extraction. After mixing an approximately equivalent weight of fresh leaves, roots, stems, and flowers, the total RNA was isolated using a Plant RNA Extraction Kit (Autolab Biotechnology, Beijing, China). Following the manufacturer's instructions, the total RNA was quantified (concentration ≥ 100 ng/μL; 28S:18S rRNA ratio ≥ 1.5) and delivered to BENAGEN (Wuhan, China) for further treatment and sequencing using the Illumina HiSeq 2500 platform. After cleaning and quality checks, high-quality (HQ) (≥ 99.5 % accuracy for single-base reads) reads were assembled using Trinity [15] with the default parameters. Contigs larger than 200 bp were used for further analysis. The longest transcript in each comparison (gene) was used as the unigene.

### Classification of gene function

A Gene Ontology (GO) enrichment analysis was performed to assign a function to each unigene. BLASTX was used to compare the unigenes with sequences in the NCBI nr protein database with an E-value cutoff of $10^{-6}$ [16]. The GO annotations of all unigenes were obtained from the Blast2GO program using the BLASTX output. Then, the GO functional classification was performed using the Web Gene Ontology Annotation Plot (WEGO) application [17].

### SSR identification

Microsatellites were identified and localized in all unigene sequences of the two *Isodon* species transcriptomes using Microsatellite (MISA, http://www.pgrc.ipk-gatersleben.de/misa). While searching for SSRs using the MISA script, the

criteria included microsatellites containing motifs between two and six nucleotides long with a minimum of six repeats for dinucleotides and five repeats for trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides. A size less than 100 bp between adjacent SSR loci indicated a compound SSR.

### DNA extraction, primer design, PCR amplification, and visualization of SSR loci

Seven *Xihuangcao* herb species were used to check the amplification of the SSR loci using newly designed primer pairs. Therefore, the total genomic DNA from the leaf tissue was isolated and purified according to a modified CTAB-based procedure [18].

The primer pairs flanking SSR regions were designed with Primer3 software (http://frodo.wi.mit.edu/primer3) using a MISA-generated Primer3 input file. The following criteria were used to design the primer pairs: product 100 - 300 bp, primer size 18 - 25 bp, melting temperature 57 - 63 °C [19], and GC content 40 – 70 %.

Three primers were used in combination for the PCR amplification. The primer set included a forward primer with an M13 sequence (5'-GTAAAACGACGGCCAGT-3') at the 5' end, a regular reverse primer, and a fluorescently labeled (FAM, HEX, TAMRA or ROX) universal M13 (-21) primer (M13: 5'-GTAAAACGACGGCCAGT-3') [20]. PCR was performed as follows: 10-µL reaction volume containing 50 ng of genomic DNA, 1 × PCR StarMix with Loading Dye (GenStar, Beijing, China), 0.0125 µM M13-tagged forward primer, 0.25 µM reverse primer, and 0.15 µM fluorescently labeled M13 primer.

The touchdown PCR program was as follows: initial denaturation for 5 min at 94 °C; 5 cycles of denaturation for 30 s at 94 °C, annealing for 30 s at 60 °C (the annealing temperature for each cycle was reduced by 1 °C per cycle) and extension for 30 s at 72 °C; 35 cycles of denaturation for 30 s at 94 °C, annealing for 20 s at 55 °C and extension for 30 s at 72 °C; and 20 min of final extension at 72 °C. The PCR products were checked on a 1.5 % agarose gel. PCR products of different lengths and fluorescence were mixed in equal proportion. The mixed PCR products were loaded onto a 3730xl DNA Analyzer (ABI, USA) for genotyping. The raw traces were analyzed using GeneMarker (SoftGenetics LLC, USA).

### Data analysis

The coefficient distances among all individual accessions were calculated using NTsys 2.10e software. Then, a dendrogram was produced based on the coefficient distances using MEGA6 [21].

## RESULTS

### De novo assembly of contigs and unigenes

Based on the two *Isodon* transcriptomes, 35,667,410 and 35,434,964 raw reads (125 bp × 2) were obtained for IS and ILL, respectively. After filtering the low-quality raw reads, 35,665,350 and 35,433,952 clean reads (high-quality reads) were used for further assembly. In total, 149,650 contigs with a mean length of 887 bp and an N50 value of 1,400 bp for IS and 103,221 contigs with an average length of 1,027 bp and an N50 value of 1,516 bp for ILL were obtained by Trinity (Table 2). The most abundant size class of contigs with lengths ranging between 200 and 800 bp was overrepresented, accounting for approximately 61.4 % of all contigs of IS and 51.1 % of all contigs of ILL. The following most abundant class included contigs between 800 bp and 2,000 bp and constituted approximately 29.0 and 36.9 % of all contigs of IS and ILL, respectively (Figure 1). The contig GC contents of IS and ILL were both approximately 44 % (Table 2). Additionally, in both datasets, the size distribution of the unigenes displayed a pattern similar to that of the contigs described above (Table 3, Figure 1).

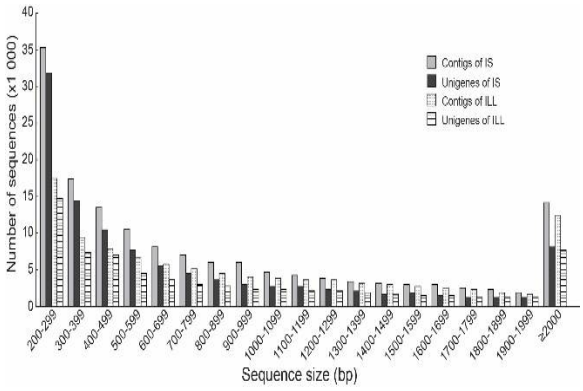


**Figure 1.** Length distribution of the contigs and unigenes in *I. serra* (IS) and *I. lophanthoides var. lophanthoides* (ILL)

### GO classification and functional annotation

The assembled sequences of IS and ILL were blasted in the NCBI nr protein database using the BLASTX algorithm with an E-value cutoff of $10^{-6}$.

**Table 2:** Summary of the transcriptomes of *I. serra* and *I. lophanthoides var. lophanthoides*

| | Species | | | |
|---|---|---|---|---|
| | ***I. serra*** | | ***I. lophanthoides var. lophanthoides*** | |
| **Feature** | *Contigs* | *Unigenes* | *Contigs* | *Unigenes* |
| Raw reads | 35,667,410 | | 35,434,964 | |
| HQ reads | 35,665,350 | | 35,433,952 | |
| Number | 149,650 | 107,777 | 103,221 | 68,220 |
| Average length (bp) | 887 | 777 | 1,027 | 966 |
| N50 (bp) | 1,400 | 1,267 | 1,516 | 1,511 |
| Maximum length (bp) | 10,202 | | 12,074 | |
| GC content | 44.17 % | 44.52 % | 44.21 % | 44.36 % |

**Table 3:** Summary statistics of the screening of the *I. serra* and *I. lophanthoides var. lophanthoides* transcriptomes for putative EST-SSRs

| **Variable** | ***I. serra*** | ***I. lophanthoides var. lophanthoides*** |
|---|---|---|
| Number of SSR-containing sequences | 11,810 (containing 14,138 SSRs) | 9,828 (containing 11,756 SSRs) |
| Number of sequences containing more than 1 SSR | 1,936 (containing 4,264 SSRs) | 1,622 (containing 3,550 SSRs) |
| Number of SSR-containing sequences used for the primer design | 6,753 (containing 7,820 SSRs) | 5,880 (containing 6,787 SSRs) |
| Number of primer pairs obtained from the primer design | 7,453 | 6,428 |

Collectively, 54.9 % (59,211/107,777) and 56.6 % (38,585/68,220) of the unigenes of IS and ILL were annotated by GO, respectively.

The GO assignments describe the gene products based on their associated cellular components, molecular functions, and biological processes. These three top categories of GO classifications were further divided into 51 functional groups (Table 4). In the “cellular component” class, the unigenes related to “cell”, “cell parts”, “organelle”, and “organelle part” were predominant. “Binding” and “catalytic activity” were the most abundant classes in the “molecular function” category. In the “biological process” category, most unigenes were involved in “cellular process”, “metabolic process”, “biological regulation”, “pigmentation”, “response to stimulus”, and “localization”. No significant difference was found between IS and ILL in any functional group.

**Identification and characterization of expressed SSRs derived from transcriptome sequences**

From these two *Isodon* transcriptomes, 6,130 of all unigenes containing SSRs (30.5 %) and 17,203 differentially expressed genes were identified, accounting for 69.5 % of all unigenes, and only 3,215 DEGs contained SSRs (accounting for 16 % of all DEGs).

In total, 14,138 SSRs were contained in 11,810 sequences (16.39 % containing more than 1 SSR locus) of IS, 11,756 SSRs were found in 9,828 sequences (16.5 % containing more than 1 SSR locus) of ILL; and the ratio was 10.96 % and 14.41 %, respectively. These SSRs included 9,370 and 8,169 dinucleotides, 4,567 and 3,461 trinucleotides, 117 and 77 tetranucleotides, 31 and 16 pentanucleotides, and 53 and 33 hexanucleotides in IS and ILL, respectively. Of the sequences containing SSRs, 6,753 and 5,880 sequences were suitable for primer design, and 7,453 and 6,428 primer pairs were ultimately obtained for IS and ILL, respectively (Table 3).

The predominant repeat motif in the IS and ILL EST-SSRs was trinucleotides (66.3 and 69.5 %, respectively), followed by dinucleotides (32.3 and 29.4 %, respectively) (Figure 3 A). Meanwhile, the proportion of tetranucleotides was only 0.8 and 0.7 % in IS and ILL, respectively, while that of both pentanucleotides and hexanucleotides was less than 0.5 % in both IS and ILL. Thus, the dinucleotide and trinucleotide repeats represented most EST-SSRs in both *Isodon* species. Dinucleotides with six repeats were the most frequent (21.1 and 24.5 %), followed by trinucleotides with five repeats (20.0 and 18.4 %), dinucleotides with seven repeats (14.5 and 16.0 %) and dinucleotides with eight repeats (12.2 and 13.0 %) in IS and ILL, respectively (Figure 3 B).

**Table 4:** Functional annotation of the assembled sequences based on Gene Ontology (GO) categorization of *I. serra* (IS) and *I. lophanthoides var. lophanthoides* (ILL)

| GO Category | Description | Unigenes (%) IS | ILL |
|---|---|---|---|
| Cellular Component | cell | 73.6 | 72.6 |
| | cell part | 73.6 | 72.6 |
| | membrane part | 26.9 | 28.2 |
| | organelle | 55.2 | 53.3 |
| | organelle part | 27.2 | 25.5 |
| | membrane | 37.4 | 38.5 |
| | macromolecular complex | 15.6 | 13.9 |
| | membrane-enclosed lumen | 6.3 | 6.0 |
| | extracellular region | 6.0 | 5.7 |
| | symplast | 4.1 | 3.9 |
| | cell junction | 4.3 | 4.1 |
| Molecular Function | catalytic activity | 51.4 | 51.4 |
| | enzyme regulator activity | 1.4 | 1.2 |
| | transporter activity | 6.9 | 6.5 |
| | binding | 65.3 | 66.2 |
| | molecular transducer activity | 2.0 | 2.3 |
| | receptor activity | 1.2 | 1.5 |
| | antioxidant activity | 0.8 | 0.7 |
| | structural molecule activity | 3.2 | 2.4 |
| | protein binding transcription factor activity | 0.5 | 0.5 |
| | nucleic acid binding transcription factor activity | 5.6 | 6.0 |
| | electron carrier activity | 0.8 | 0.7 |
| Biological Process | developmental process | 14.2 | 14.6 |
| | cellular process | 64.7 | 64.3 |
| | single-organism process | 51.8 | 52.2 |
| | multicellular organismal process | 14.1 | 14.8 |
| | cellular component organization or biogenesis | 12.6 | 13.1 |
| | localization | 15.1 | 14.4 |
| | biological regulation | 25.3 | 25.4 |
| | establishment of localization | 14.4 | 13.7 |
| | signaling | 8.8 | 8.8 |
| | regulation of biological process | 23.7 | 23.9 |
| | immune system process | 3.2 | 4.1 |
| | metabolic process | 61.0 | 59.8 |
| | multi-organism process | 7.1 | 7.2 |
| | positive regulation of biological process | 4.1 | 4.1 |
| | response to stimulus | 28.0 | 29.2 |
| | reproductive process | 6.9 | 7.2 |
| | reproduction | 7.5 | 8.0 |
| | negative regulation of biological process | 4.6 | 4.7 |
| | growth | 2.9 | 3.0 |

The length of the SSR region (motif length × repeat number) ranged from 12 to 66 bases, and 12 bases was the most frequent number (21.1 % and 24.5 %), followed by 15 bases (20.0 and 18.4 %) and 18 bases (17.9 and 16.3 %) (Figure 4B) in IS and ILL, respectively. The number of repeats in the different SSR motifs ranged from 5 to 20. Six repeats and five repeats were the most frequent numbers (Figure 3C). Among the dinucleotide repeats, AG/CT was the most common motif (68.6 and 72.8 %), followed by AC/GT (18.3 and 16.9 %), AT/AT (13.1 and 10.2 %), and CG/CG (0.03 and 0.2 %) in IS and ILL, respectively. Among the trinucleotide repeats, AAG/CTT was the most frequent, accounting for 19.9 and 18.2 % in IS and ILL, respectively. The numbers of tetranucleotides, pentanucleotides and hexanucleotides were too small for statistical analysis.

## Validation of SSR markers and genetic identification of 'Xihuangcao' herbs

A subset of 100 EST-SSR markers from the IS dataset was selected for validation. Fifty of these markers were located in terpene synthesis genes, and the other markers were randomly chosen. DNA extracted from seven 'Xihuangcao' herb samples (Table 1) was used as the PCR template. PCR products with a single bright band

were genotyped using a 3730xl DNA Analyzer (Applied Biosystems, Carlsbad, CA, USA) (Figure 4).

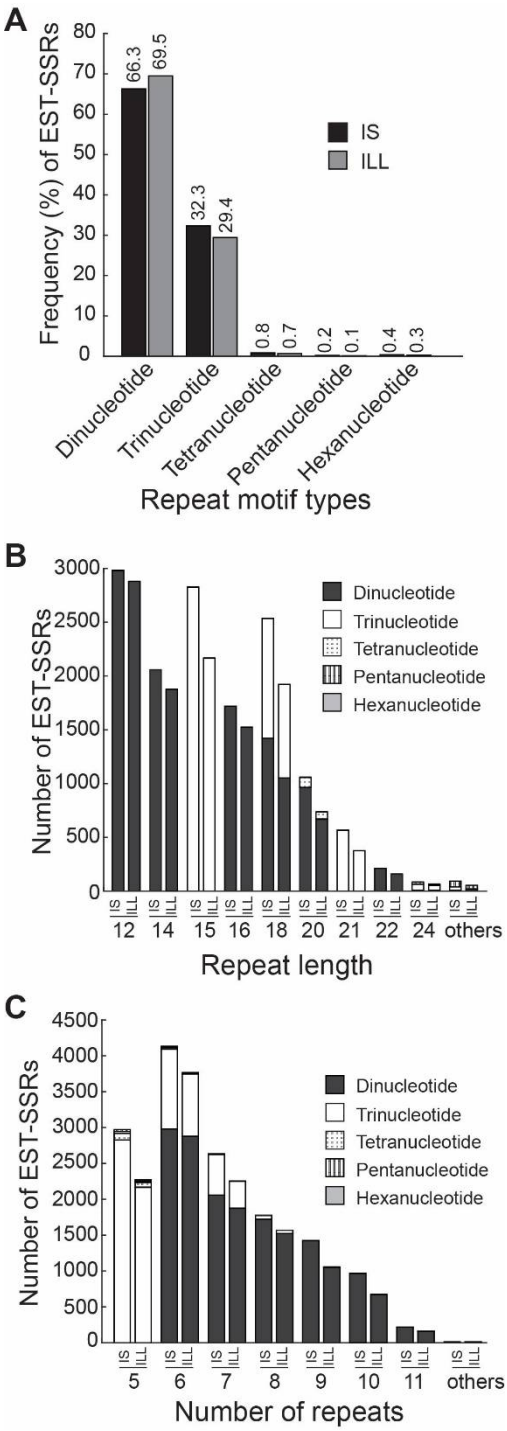


**Figure 3:** Characteristics of SSR markers in two *Isodon* species. A: Frequency distribution of EST-SSRs based on the type of repeat motif. B: Frequency distribution of EST-SSRs based on the repeat length (bars without and with white points represent *I. serra* [IS] and *I. lophanthoides var. lophanthoides* [ILL], respectively). C: Number of repeats of different types of SSR motifs (bars without and with white points represent *I. serra* and *I. lophanthoides var. lophanthoides*, respectively)

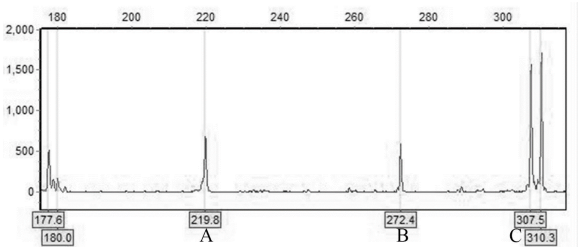


**Figure 4:** Analysis (GeneMapper output) of 3 SSR markers of "Xihuangcao" herbs in a single capillary: *I. lophanthoides* is used as an example. A, B, and C show the peak waves of the amplification products using the SSR22, SSR21, and SSR04 primers, respectively

Of the 100 SSRs chosen for validation, 53 primer pairs were suitable as molecular markers. To ensure the accuracy of the results, 24 primer pairs with bright bands, higher polymorphism, and single amplification were selected as the core primer set (Tables 5 and 6). The genotyping data obtained using these 24 core primer pairs were clustered, analyzed and verified using NTSYS 2.10e software (Figure 5).

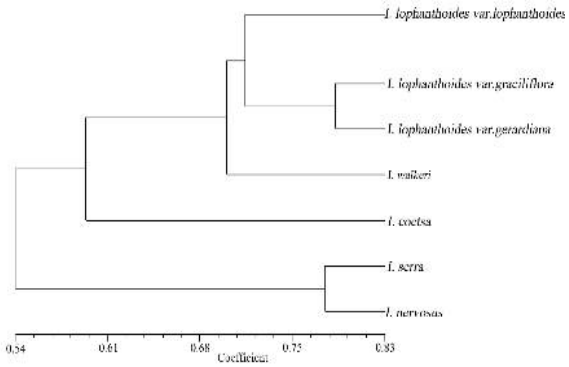


**Figure 5:** Genetic relationship among seven accessions of *Xihuangcao* herbs based on a neighbor-joining tree constructed using genotypic data from 53 or 24 polymorphic EST-SSRs

The genetic similarity coefficient variation ranged from 0.54 to 0.83. The seven germplasm resources of the *Xihuangcao* herbs could be divided by 0.65. According to the phylogenetic diagram, these seven species of *Xihuangcao* herbs could be divided into the following two groups: Group I included IS and *I. nervosus*, which were closely related, and Group II included *I. lophanthoides*, *I. lophanthoides var. graciliflora*, *I. lophanthoides var. gerardiana*, *I. coetsa* and *I. walkeri*. The relationship among *I. lophanthoides var. gerardiana*, *I. lophanthoides var. graciliflora* and *I. lophanthoides* was the closest, followed by *I. walkeri* and then *I. coetsa*. These results indicate the potential utility of these markers in the genetic identification of *Isodon* species.

**Table 5**: The 24 EST-SSR primer pairs

| SSRID | ID | SSR | Upstream primer (5'-3') | Downstream primer (5'-3') |
|---|---|---|---|---|
| SSR01 | c46903_g2 | (GGT)6 | GTGCACTTTTCCGCTTTCTC | GTGAAATCCCCACAAACCAC |
| SSR02 | c35340_g1 | (CA)8 | CCATTTCTCTAGCCCTCTCTCA | TCCACCCTCAATAGCCTGTC |
| SSR03 | c4910_g1 | (CT)8 | CCGGTCGTTAAGATTTGCAT | GTTGTGCATTTATTGCGGTG |
| SSR04 | c49288_g3 | (AT)6 | CAGAAAACCAAGACTCTCTTCCA | ACGGCTTGTTCAACCTCATC |
| SSR05 | c43304_g1 | (AG)8 | TTATACCCTCCCTTTTCCCG | GAGGCATATGACTGGGGAGA |
| SSR06 | c10552_g1 | (CCT)5 | CCACCGTCTAACCACCAAGT | GGAGGAAGAGGAGAGGAGGA |
| SSR07 | c38502_g1 | (TC)7 | CCACCATTGAAATCAATCCC | CTCACACTCAACACACACTCATC |
| SSR08 | c7870_g1 | (TC)9 | AACCCCATTTTTCCTCAACC | AGAGGGTGGGAGAACAGGAT |
| SSR09 | c88702_g1 | (CT)7 | GAAGGCCTGATCGTTCTCCT | AAGCGGCTGTTGCTTCTTTA |
| SSR10 | c24826_g1 | (AT)6 | CTGGGAAATTGTGGCTTCAT | GCGCTATAACCGAAAGACATT |
| SSR11 | c37352_g2 | (AG)6 | GGGTCTTTCCAAGAGAAGGG | TTTTCTATGTCGGCCTCAGTC |
| SSR12 | c48440_g1 | (TG)6 | TGTTGTGCGTGATGGAATTT | ATATGCCCCATTTGCTTTTG |
| SSR13 | c43529_g1 | (AGC)6 | CCTCTTTTGAAATTGGAGCG | ACCCTCGGTTACCCCATTAC |
| SSR14 | c47319_g1 | (GCG)6 | GAGGCGTTCAAGAAGTACGC | GTGAACTTGGCAATGTGGTG |
| SSR15 | c2672_g1 | (TGA)10 | AACCACAACAGAAACAGCCC | ACATGGAGCTAGGCAAGCAT |
| SSR16 | c39603_g1 | (AGG)8 | AAGATCCGAGGAGAGCAACA | GTGAGGGGAGGGATAGGGTA |
| SSR17 | c5937_g1 | (ATT)8 | ATTGAAGTTGGTGCCCTGAG | CACCCTAATCACAATTACCAACAC |
| SSR18 | c48051_g3 | (ATC)8 | CCAGACAGAACTCCTCCAGC | AACACGTGGAGAAATCGAGG |
| SSR19 | c36658_g2 | (ATC)8 | CTAGGGCATGTAGGTTGGGA | CTGAAGAGCAACGACGATGA |
| SSR20 | c48051_g1 | (CGC)8 | CTCCTCGACAACAGCAACAA | CCATTCCCTGATCTTCCTCA |
| SSR21 | c31063_g1 | (GAG)8 | GAGCCAGAGGTGGAGTTGAC | AAAATCATCCCTCCCCAATC |
| SSR22 | c45967_g3 | (AAC)8 | GGTTTTAGGCTTAGGGTGGC | AGGAGATTAACGCAGCGAGA |
| SSR23 | c22078_g2 | (TCA)8 | TACTCTCCCAGTCCGGTGAT | GTATCCTTGGGTCGACATGG |
| SSR24 | c13707_g1 | (ACC)8 | TGGTGCTGTACATGTTGCCT | ATTCCTTCGACTGGATGGTG |

## DISCUSSION

Recently, large-scale transcriptome sequencing of many species using Illumina paired-end sequencing technology has been performed. This technique is cost-effective and time-saving compared to conventional sequencing methods. For example, the Illumina HiSeq 2500 system can generate up to 1 terabase (Tb) of sequencing data in less than six days (http://www.illumina.com/). Thus, this approach could provide an innovative way to expedite the analysis of the few studied species [e.g., herbal plants [22] and endangered species [23]]. A limitation of Illumina platforms is the short read length (less than 150 bases) compared to Roche 454 (up to 1,000 bases) (http://www.454.com/). The assembly of short reads may present significant challenges. Currently, Illumina reads can be assembled using several sequence assembly tools designed for very short reads, e.g., Velvet, AbySS, and Trinity [24]. Trinity is widely considered the best de novo RNA-Seq assembler [15].

The average contig lengths of IS and ILL in this study were 777 bp and 966 bp, respectively, which are comparable to those in other studies *Capsicum frutescens*, 729 bp [5] and *Petunia hybrid,* 822 bp [25]) using Trinity as the assembler. Longer raw reads (125 bp) provided noticeably better results in further gene functional analysis and molecular marker development.

**Table 6:** Fluorescence genotyping results of multiple SSRs in 7 species/varieties of "Xihuangcao" herbs

| SSR ID | A (bp) | B (bp) | C (bp) | D (bp) | E (bp) | F (bp) | G (bp) |
|---|---|---|---|---|---|---|---|
| SSR01 | 259/288 | 266/284 | 288 | 288 | 266/285 | 256/284 | 238/245 |
| SSR02 | 254 | 254 | 254/268 | 268 | 254 | 254 | 266 |
| SSR03 | 235/245 | 211 | 245 | 245/247 | 211 | 245 | 241 |
| SSR04 | 284 | 285 | 287 | 286 | 285 | 283 | 282 |
| SSR05 | 285 | 248 | 244/248 | 244/248 | 248 | 244 | 248 |
| SSR06 | 283 | 283/285 | 284/287 | 283/289 | 284 | 283/285 | 282 |
| SSR07 | 297 | 290 | 297 | 297 | 284 | 282 | 279 |
| SSR08 | 254/273 | 255/274 | 254 | 254/273 | 255/274 | 254/273 | 254 |
| SSR09 | 136/164 | 178 | 179/189 | 179 | 178 | 178 | 174 |
| SSR10 | 247 | 247/250 | 246/250 | 253 | 245/251 | 246/250 | 247/249 |
| SSR11 | 297/301 | 283/301 | 301/304 | 298/301 | 301 | 300 | 297 |
| SSR12 | 215 | 183 | 216 | 216 | 194 | 195 | 193/197 |
| SSR13 | 243/248 | 252 | 242/248 | 242/248 | 252/254 | 242/248 | 263 |
| SSR14 | 301 | 304 | 301 | 301 | 304 | 301 | 303 |
| SSR15 | 240 | 244/259 | 241 | 240 | 240 | 240 | 249/267 |
| SSR16 | 304 | 301 | 304 | 304 | 301 | 304 | 305 |
| SSR17 | 233 | 197/236 | 239 | 217/239 | 197/205 | 238 | 223/236 |
| SSR18 | 272 | 272 | 260/272 | 258 | 264 | 260/272 | 260 |
| SSR19 | 220 | 220/222 | 220 | 220 | 219/222 | 220 | 218 |
| SSR20 | 260 | 261/271 | 272 | 262 | 261/271 | 272 | 272 |
| SSR21 | 307/310 | 301/303 | 310 | 311 | 301/303 | 310 | — |
| SSR22 | 295/308 | 285/305 | 309 | 309 | 285/305 | 309 | 286/310 |
| SSR23 | 198 | 206/221/237 | 197 | 198/220 | 206/237 | 205/222 | 197 |
| SSR24 | 317 | 308 | 325 | 321 | 325 | 317/325 | — |

*Note: "—" indicates that there were no corresponding amplification products*

As described in previous studies, the overall frequency and the frequency of the different lengths of EST-SSRs and repeat motifs are significantly influenced by the size of the dataset, the percentage of repetitive DNA, and the search criteria [26]. In the present study, 107,777 and 68,220 unique sequences (approximately 83.8 Mb for IS and 65.9 Mb for ILL) were used to search for SSRs, resulting in 1,726 and 2,366 SSRs derived from 1,583 (3.9 %) and 2,189 (4.3 %) SSR-containing sequences in IS and ILL, respectively. The SSR abundance in Isodon was higher than that in flax (3.5 %) [27], wheat (3.6 %)[28],, and lower than that in sesame (8.9 %) [29].

The overall frequency of EST-SSRs was 1/14.1 kb and 1/12.7 kb in IS and ILL, respectively, in the present investigation, which was lower than that in coffee (1/1.6 kb) [30], sesame (1/7.0 kb) [29], and bread wheat (1/9.2 kb) [31] and higher than that in wheat (1/15.6 kb) [28] and flax (1/16.5 kb) [27], indicating that *Isodon* has a moderate general SSR frequency.

In IS and ILL, dinucleotides (55.7 and 52.1 %) were present at the highest frequency of all repeat motifs, followed by trinucleotides (42.5 and 45.3 %). This finding is consistent with results obtained in sesame [29], coffee [30], and lotus [32]. In contrast, trinucleotides were the most abundant class in radish [33], and *Cucurbita pepo* [34].

The relative abundance of dinucleotides and trinucleotides is strongly influenced by the detection parameters and characteristics of the EST database analyzed in different species. In the present study, the lowest number of repeats of dinucleotides and trinucleotides was six and five, respectively. When the minimum numbers of repeats of dinucleotides and trinucleotides were both five, the percentages of dinucleotides and trinucleotides were 61.7 and 37.1 % in IS and 63.4 and 34.8 % in ILL, respectively. Thus, the search criteria used for SSR mining can substantially affect the observed differences.

AG/CT was the predominant repeat motif among the dinucleotides in both IS and ILL (48.6 and 47.9 %, respectively), which is consistent with the results reported in similar studies [27,32,33]. CG/CG was the least abundant in both species (0.3 and 0.5 % in IS and ILL, respectively), which has also been reported in other plant species, such as sesame [29] and coffee [30]. No CG/CG repeat motif has been found in rice, corn, soybean [35], flax [27], or wheat [36]. The most frequent trinucleotide motif was AAG/CTT in IS and ILL (27.3 and 28.9 %, respectively), which is consistent with previous reports investigating flax [27], lotus [31], radish [32], coffee [30] and *Cucurbita pepo* [33].

Thus far, no SSR markers have been developed for Isodon, limiting population structure analyses,

QTL/gene mapping, and marker-assisted selection. Therefore, the development of more SSR markers for Isodon is urgently needed. In the present study, 7,453 and 6,428 putative SSR markers were developed for IS and ILL, respectively. This study is the first to report the development of SSR markers for Isodon.

'Xihuangcao' cultivars are easily confused, and the current identification standards based on the traditional classification remain controversial. Some experts and scholars refer to *I. lophanthoides* var. *gerardiana* (Benth.) Hara as *I. lophanthoides* var. *graciliflora* (Benth.) Hara on the market. In addition, *I. lophanthoides* recorded in "The Compilation of Chinese Herbal Medicine" is referred to as *I. lophanthoides* var. *graciliflora* (Benth.) Hara [37]. In the present experiment, the results were almost the same as those reported by Mo et al. [13], who used RAPD to identify the 4 'Xihuangcao' resources. The results of the present study also demonstrate that the presented EST-SSR primer pairs could practically, rapidly, and accurately identify the common varieties of 'Xihuangcao' herbs.

## CONCLUSION

In this study, the transcriptomes of two *Isodon* species were characterized, and an unprecedented number of genomic resources were obtained for these important medicinal plants. Numerous EST-SSR markers were developed for 'Xihuangcao' herb, and 24 primer pairs were selected for identification research. There is great potential for future research involving genetic mapping, cloning, and molecular marker-assisted selection of 'Xihuangcao' herb. Studying the evolutionary relationships between wild-type species and cultivated varieties of 'Xihuangcao' herb could provide a reference for selecting breeding parents.

## DECLARATIONS

### *Acknowledgement*

The authors appreciate the funding provided by the Guangdong Provincial Key Laboratory of Applied Botany, the South China Botanical Garden; the Chinese Academy of Sciences; the Science and Technology Project of Guangdong Province (nos. 2015A040404029 and 2016B020239004), and the Project of the Bureau of Traditional Chinese Medicine of Guangdong Province (no. 20151027). We also thank the Key Laboratory of Plant Resources Conservation and Sustainable Utilization for providing the laboratory equipment and venues.

### *Conflict of interest*

No conflict of interest is associated with this work.

### *Contribution of authors*

We declare that this work was performed by the authors named in this article and that all liabilities pertaining to claims related to the content of this article will be borne by the authors. Shanshan Huang and Weiming Hu contributed equally to this paper. Shanshan Huang designed the study and performed the experiments; Shanshan Huang and Weiming Hu analyzed the data and wrote the manuscript. Shaohua Zeng, Xiaolu Mo and Ying Wang provided advice.

## REFERENCES

1. *Sun HD, Huang SX, Han QB. Diterpenoids from Isodon species and their biological activities. Nat Prod Rep 2006; 23(5): 673-698.*
2. *Feng WS, Zang XY, Zheng XK, Wang YZ. A new phenylethanoid glycoside from Rabdosia lophanthoides (Buch.-Ham.ex D.Don) Hara. Chin Chem Lett 2009; 20(4): 453-455.*
3. *Li H-W. Taxonomic review of Isodon (Labiatae). J Arnold Arbor 1988; 69(4): 289-400.*
4. *Chen LJ, Miao Y, Lai XP. Pollen morphology of Chinese medicine Xihuangcao and its related species. Journal-Xiamen University Natural Science 2000; 39: 549-554.*
5. *Liu S, Li W, Wu Y, Chen C, Lei J. De novo transcriptome assembly in chili pepper (capsicum frutescens) to identify genes involved in the biosynthesis of capsaicinoids. PloS One 2013; 8: e48156.*
6. *Huang SX, Xiao WL, Li LM, Li SH, Zhou Y, Ding LS, Lou LG, Sun HD. Bisrubescensins A-C: three new dimeric ent-kauranoids isolated from Isodon rubescens. Org Lett 2006; 8(6): 1157-1160.*
7. *Lin L, Dong Y, Yang B, Zhao M. Chemical constituents and biological activity of Chinese medicinal herb 'Xihuangcao'. Comb Chem High Throughput Screen 2011; 14(8): 720-729.*
8. *Lin L, Zhuang M, Lei F, Yang B, Zhao M. GC/MS analysis of volatiles obtained by headspace solid-phase microextraction and simultaneous-distillation extraction from Rabdosia serra (MAXIM.) HARA leaf and stem. Food Chem 2013; 136(2): 555-562.*
9. *Chen C, Chen Y, Zhu H, Xiao Y, Zhang X, Zhao J, Chen Y. Effective compounds screening from Rabdosia serra (Maxim) Hara against HBV and tumor in vitro. Int J Clin Exp Med 2014; 7(2): 384-392.*
10. *Liu YH, Huang SX, Zhao QS, Ding JK, Goh NK, Sun HD, Chia TF, Tan SN, Chia LS. A new ent-kauranoid from*

*Isodon lophanthoides var. geradianus. Nat Prod Res 2008; 22(10): 860-864.*

11. *Lai X, Chen L, Chen J, Li Z. Micromorphological identification of leaves of the botanical origins of Xihuangcao in Guangdong. Journal of Guangzhou University of Traditional Chinese Med 1995; 13(3/4): 83-85.*
12. *Chen J, Jiang D, Zhao A, Lai X. Microscopic identification of medicinal material of herba rabdosiae serrae. Journal of Guangzhou University of Traditional Chinese Medicine 2006; 1:17.*
13. *Mo XL, Zeng QQ, Huang SS, Cai YW, Wang YS, Yan Z. Identification of herba rabdosiae serrae from different plant resources by RAPD method. Zhong Yao Cai 2012; 35(9): 1388-1391.*
14. *Chen L, Chen Y, Qu L, Ye C, Lai X. RAPD analysis on Chinese medicine Xihuangcao and its related species. Acta Scientiarum Naturalium Universitatis Sunyatseni 1998; 38: 102-106.*
15. *Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013; 8(8): 1494-1512.*
16. *Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25(17): 3389-3402.*
17. *Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 2006; 34: W293-W297.*
18. *Mace ES, Buhariwalla KK, Buhariwalla HK, Crouch JH. A high-throughput DNA extraction protocol for tropical molecular breeding programs. Plant Mol Biol Report 2003; 21(4): 459-460.*
19. *Yadav HKM, Ranjan A, Asif MH, Mantri S, Sawant SV, Tuli R. EST-derived SSR markers in Jatropha curcas L.: development, characterization, polymorphism, and transferability across the species/genera. Tree Genet Genomes 2011; 7(1): 207-219.*
20. *Schuelke M. An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 2000; 18(2): 233-234.*
21. *Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013; 30(12):2725-2729.*
22. *Lulin H, Xiao Y, Pei S, Wen T, Shangqin H. The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PLoS One 2012; 7(6): e38653.*
23. *Hao da C, Ge G, Xiao P, Zhang Y, Yang L. The first insight into the tissue specific taxus transcriptome via illumina second generation sequencing. PLoS One 2011; 6(6): e21220.*
24. *Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 2011; 12 Suppl 14: S2.*
25. *Villarino GH, Bombarely A, Giovannoni JJ, Scanlon MJ, Mattson NS. Transcriptomic analysis of Petunia hybrida in response to salt stress using high throughput RNA sequencing. PLoS One 2014; 9(4): e94651.*
26. *Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol 2005; 23(1): 48-55.*
27. *Cloutier S, Niu Z, Datla R and Duguid S. Development and analysis of EST-SSRs for flax (Linum usitatissimum L.). Theor Appl Genet 2009; 119: 53–63.*
28. *Kantety RV, La Rota M, Matthews DE, Sorrells ME. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 2002; 48(5-6): 501-510.*
29. *Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X. Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics 2011; 12: 451.*
30. *Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor Appl Genet 2007; 114(2): 359-372.*
31. *Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomics 2003; 270: 315-323.*
32. *Pan L, Xia Q, Quan Z, Liu H, Ke W, Ding Y. Development of novel EST-SSRs from sacred lotus (Nelumbo nucifera Gaertn) and their utilization for the genetic diversity analysis of N. nucifera. J Hered 2010; 101(1): 71-82.*
33. *Wang S, Wang X, He Q, Liu X, Xu W, Li L, Gao J, Wang F. Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. Plant Cell Rep 2012; 31: 1437-1447.*
34. *Blanca J, Cañizares J, Roig C, Ziarsolo P, Nuez F, Picó B. Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (Cucurbitaceae). BMC Genomics 2011; 12:104.*
35. *Gao L, Tang J, Li H, Jia J. Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed 2003; 12(3): 245-261.*
36. *Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy P, Bernard M, Sourdille P. Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet 2004; 109(4): 800-805.*
37. *Yan F-L, Xie R-J, Yin Y-Y, Zhang Q. Serrin D, a new ent-kaurane diterpenoid from Isodon serra. J Chem Res 2012; 36(9): 523-524.*