西北植物学报，2019，39(1)：0059－0067
Acta Bot. Boreal. -Occident. Sin.

doi：10.7606/j.issn.1000-4025.2019.01.0059 http：//xbzwxb.alljournal.net

基于转录组测序的葛根 SSR 标记研究与利用

肖 亮[1,2]，尚小红[1,2]，曹 升[1,2]，谢向誉[1,2]，曾文丹[1,2]，严华兵[1*]

（1 广西农业科学院 经济作物研究所，南宁 530007；2 广西作物遗传改良生物技术重点开放实验室，南宁 530007）

摘　要：该研究以'桂粉葛 1 号'为材料，通过转录组测序的方法测得 8.9 Gb clean reads，组装成 137 629 个转录本，最终得到 83 811 个 Unigene 序列。进化树分析表明，葛根和苜蓿、花生聚为一支。用 MISA 软件在 83 811 个序列中检测到 25 452 个简单重复序列（SSR）位点，三核苷酸重复的 SSR 数量最多，其次是二核苷酸重复。三核苷酸重复中，$(AAG)_n$ 是最普遍的重复单元（27.87％）。共设计了 229 对 SSR 引物，其中 28 对引物可以产生清晰的条带和丰富的多态性，被用于检测 44 份葛根资源的遗传多样性。在 44 个葛根资源的基因组 DNA 中共扩增出 90 个片段，其中 89 个条带有多态性，平均等位基因数为 3.178 6。多态性信息量范围为 0.083 0～0.774 2（平均数为 0.455 7）。聚类分析显示遗传相似性系数范围为 0.266 7～1.000 0。这些结果提示所检测的葛根资源在 DNA 分子水平存在着丰富的遗传多样性。当阈值为 0.58 时，44 个资源可以划分为 2 个类群，且 44 份资源的类群划分与地理来源之间没有直接关系，但这些标记将是葛根遗传多样性研究可用的基因组资源。

关键词：葛；SSR 标记；资源收集；遗传多样性

中图分类号：Q346^{+}.5；Q789　　　　**文献标志码**：A

Utilization of Simple Sequence Repeat（SSR）Markers Developed from a *de novo* Transcriptome Assembly in *Pueraria thomsonii* Benth.

XIAO Liang[1,2]，SHANG Xiaohong[1,2]，CAO Sheng[1,2]，XIE Xiangyu[1,2]，
ZENG Wendan[1,2]，YAN Huabing[1*]

（1 Cash Crops Research Institute，Guangxi Academy of Agricultural Sciences，Nanning 530007，China；2 Guangxi Crop Genetic Improvement and Biotechnology Key Lab，Nanning 530007，China）

Abstract：In this study，137 629 transcripts were assembled from 8.9 Gb of clean Illumina DNA sequencing read data，yielding 83 811 unigene sequences from *Pueraria thomsonii* 'No. 1'. A phylogenetic analysis indicated that *Pueraria lobata* and *Medicago sativa* clustered together with *Arachis hypogaea*（peanut）. We detected 25 452 SSR loci in the 83 811 assembled unigenes using MISA software. Tri-nucleotide repeats were the most abundant followed by di-nucleotide repeats. Among the tri-nucleotide motifs，$(AAG)_n$（27.87％），was the most common repeat unit. A total of 229 SSR primer pairs were designed，and 28 markers that gave clear，polymorphic amplification products were used to analyze the genetic diversity within a panel of 44 *Pueraria* accessions. Ninety SSR fragments，consisting of 89 alleles，were amplified from genomic DNA of the 44 accessions. The average allele number is 3.178 6. Polymorphic information content（PIC）values ranged between 0.083 0 and 0.774 2（mean ＝ 0.455 7）. Cluster analysis showed that the genetic similarity coefficients among the accessions ranged from 0.266 7 to 1.000 0. These results

收稿日期：2018-11-08；**修改稿收到日期**：2019-01-17
基金项目：广西科技计划工程（桂科 AD17195072）；广西科技重要工程（桂科 AA17202027-4）；广西自然科学基金青年基金（2018GXNSFBA294001）
作者简介：肖　亮（1982－ ），男，博士，副研究员，主要从事葛根、木薯的分子生物学研究。E-mail：xiaoliang0918@163.com
＊**通信作者**：严华兵，研究员，主要从事木薯、葛根等无性繁殖作物的遗传育种研究。E-mail：h.b.yan@hotmail.om

suggest that the detected *P. lobata* resources have abundant genetic diversity at DNA molecular level. The in-group similarity coefficient (0.58) was observed in the 44 germplasm accessions, and all accessions could be clearly divided into two groups. The clustering results of tested *P. lobata* resource did not show clear correlation to their geographic origin. These markers are reliable genomic resource for genetic diversity analysis in *Pueraria*.

Key words: *Pueraria*; SSR markers; germplasm collection; genetic diversity

Pueraria DC., a perennial vine that is sometimes commonly known as kudzu, has been used in traditional Chinese medicine for centuries. *Pueraria thomsonii* Benth. and *Pueraria lobata* (Willd.) Ohwi are two important species of *Pueraria*. The main components of *P. lobata* roots consist of starch, cellulose, and isoflavonoids. Puerarin and daidzin, two isoflavonoids found in *P. lobata*, have been used for the prevention and treatment of cardiovascular disease, hypolipidemia, angina pectoris, and diabetes[1-2], and have a significant effect on increasing blood flow in the coronary artery and regulating blood circulation[3].

China is the center of distribution of *Pueraria*, with a long history of growing *Pueraria* species. Unfortunately, some excellent *Pueraria* germplasm has become rare or has disappeared due to excessive mining and the lack of resource protection. It is urgent to analyze the genetic diversity in order to protect rare allelic variation in *P. lobata*. Molecular markers can be used to explore the diversity of germplasm resources and the relationship among and between varieties[4]. Most recently, RAPD (random amplified polymorphic DNA), ISSR (inter-simple sequence repeat) and SRAP (sequence-related amplified polymorphic) markers have been used to analyze the genetic diversity in *P. lobata*[5-9].

Microsatellite (also known as simple sequence repeats, SSRs) markers are co-dominant, abundant, and multi-allelic, and the SSR motifs are uniformly distributed over the genome[10]. However, the isolation of genomic SSR markers from genomic DNA libraries is time consuming and requires a considerable financial investment. Genic-SSRs have the advantages of expressed sequence tag (EST) SSRs, but contain more comprehensive information than do EST-SSRs.

In this study, we present the transcriptome of the species *P. thomsonii* Benth. The objectives of this study were to i) develop genic-SSR markers based on RNA-seq data; ⅱ) construct a phylogenetic tree of *P. thomsonii* accessions; ⅲ) identify some polymorphic markers in a panel consisting of 44 *Pueraria* germplasm collections from the Guangxi area in southern China.

1 Materials and methods

1.1 Plant materials

The genotype used for transcriptome sequencing, *P. thomsonii* Benth. 'No. 1', was cultivated on the Guangxi Academy of Agriculture Sciences (GXAAS) farm. Root, leaf, and stem tissue from three-month-old *Pueraria* plants were collected and flash frozen in liquid nitrogen and stored at −80 ℃. All 44 germplasm collections, including 40 *P. lobata* and four *P. thomsonii*, were collected from the Guangxi area (Table 1).

1.2 Methods

1.2.1 RNA extraction and sequencing Total RNA was isolated using Trizol reagent (Invitrogen, CA, USA). RNA integrity was assessed using the RNA Nano 6000 Assay Kit on the Agilent Bioanalyzer 2 100 system. A total of 1.5 g RNA per sample was used for cDNA library construction. The transcriptome libraries were constructed according to Zhang *et al.*[11]. The library preparations were sequenced on an Illumina Hiseq 4 000 platform, and 150 bp paired-end reads were generated. Clean reads were obtained by removing reads containing adapters, reads containing runs of poly-Ns (unknown bases), and low quality reads from the raw data. Transcriptome assembly was performed with Trinity software[12] using the clean,

Table 1　*Pueraria* germplasm collections used in this study and their geographical origins in China

No.	Species	Origin	No.	Species	Origin
1	*P. lobata*	Jinchengjiang，Hechi	23	*P. lobata*	Wuming，Nanning
2	*P. lobata*	Tianyang，Baise	24	*P. lobata*	Guangxi
3	*P. lobata*	Duan，Hechi	25	*P. lobata*	Guangxi
4	*P. lobata*	Pingle，Guilin	26	*P. lobata*	Yanshan，Guilin
5	*P. lobata*	Lipu，Guilin	27	*P. lobata*	Laoshan，Baise
6	*P. lobata*	Zhongshan	28	*P. lobata*	Laoshan，Baise
7	*P. lobata*	Tianlin，Baise	29	*P. lobata*	Laoshan，Baise
8	*P. lobata*	Tianlin，Baise	30	*P. lobata*	Huixian，Guilin
9	*P. lobata*	Beiliu，Yulin	31	*P. lobata*	Huixian，Guilin
10	*P. thomsonii*	Tengxian，Wuzhou	32	*P. lobata*	Huixian，Guilin
11	*P. lobata*	Yangsuo，Guilin	33	*P. lobata*	Pingle，Guilin
12	*P. lobata*	Tengxian，Wuzhou	34	*P. lobata*	Longan，Nanning
13	*P. thomsonii*	Tengxian，Wuzhou	35	*P. lobata*	Dongxin，Fangchenggang
14	*P. lobata*	Cenxi，Wuzhou	36	*P. lobata*	Fangchenggang
15	*P. lobata*	Wuming，Nanning	37	*P. lobata*	Pingle，Guilin
16	*P. lobata*	Pingle，Guilin	38	*P. lobata*	Rongshui，Liuzhou
17	*P. lobata*	Jinxiu，Laibing	39	*P. lobata*	Luocheng，Hechi
18	*P. thomsonii*	Tengxian，Wuzhou	40	*P. thomsonii*	Nanning
19	*P. lobata*	Tengxian，Wuzhou	41	*P. lobata*	Tengxian，Wuzhou
20	*P. lobata*	Bobai，Yulin	42	*P. lobata*	Tengxian，Wuzhou
21	*P. lobata*	Guangxi	43	*P. lobata*	Pingle，Guilin
22	*P. lobata*	Wuming，Nanning	44	*P. lobata*	Pingle，Guilin

high quality reads. To analyze the transcriptome without a reference genome，the sequences were assembled into transcripts，which could then be hierarchically clustered with the Corset program[13]. After Corset clustering，the longest cluster sequence is chosen from a group of isoforms generated for further analysis.

1.2.2　Evolutionary analysis of protein sequences

The coding sequences（CDS）were detected. The genomic and GTF（gene transfer format）sequences from nine species，alfalfa（*Medicago sativa*），peanut（*Arachis hypogaea*），and *Arabidopsis et al*，were downloaded from the Ensemble database，and the corresponding CDS were obtained for comparative phylogenetic analysis.

Predicted protein sequences of the 170 shared single-copy genes were aligned using Muscle 3.8.31[14]. OrthoMCL was used to search the orthologous genes for multiple sequence alignment results[15]. Predicted protein sequences of the 170 shared single-copy genes were aligned using Muscle3.8.31[14]. The phylogenetic analysis was conducted using the program MEGA7[16].

1.2.3　SSR discovery and PCR primer design

SSRs were identified in the transcriptome using the Perl script MIcroSAtellite（MISA）（http://pgrc.ipk-gatersleben.de/misa/misa.html）. SSRs consisting of repeat units of 2－6 nucleotides were considered for further development. The minimum SSR length criteria were defined as five reiterations of each repeat unit. Di-nucleotide motifs were retained only when the numbers of repeats were no less than ten and six，respectively. Primer pairs for each SSR were designed using Primer3 software（http://primer3.source.net/releases.php）. The criteria for primer design were as follows：length of 18－23 bases with an optimum of 20 bases；a primer annealing temperature range of 52－60 ℃

with an optimum of 55 ℃; and (G+C) content of 30%—70% with an optimum of 50%.

1.2.4 Plant genomic DNA extraction and SSR genotyping Genomic DNA was extracted from leaf tissue of all 44 genotypes using the CTAB method[17]. All SSR markers were amplified in six varieties in order to select the optimal markers. The PCR reactions were performed in 20 μL volumes containing 2 μL DNA (50 ng/μL), 10 μL 2X TAQ Mix (0.5 U *Taq* DNA polymerase per μL), 1 μL left primer, 1 μL right primer, and 6 μL sterile dd H_2O. The PCR cycling conditions were: initial denaturation at 94 ℃ for 5 min, followed by 30 cycles of 30 s at 94 ℃, 30 s at 55 ℃, and 30 s at 72 ℃, with a final extension for 10 min at 72 ℃. The PCR products were examined by electrophoresis on 6% non-denaturing PAGE gels and silver stained as described by Zhang *et al*.[18].

1.2.5 Genetic diversity and cluster analysis The highly polymorphic SSR markers were used to genotype the 44 *Pueraria* germplasm accessions. The genetic diversity parameters were established using POWERMARKER V3.25, and the NTSYS-PC program was used to evaluate the genetic relationships among the 44 accessions.

2 Results and analysis

2.1 Assembly of transcriptome contigs in *P. thomsonii* Benth.

A cDNA library was constructed from RNA extracted from *P. thomsonii* 'No. 1' and sequenced on an Illumina Hiseq 4 000 instrument, yielding a total of 9.2 Gb raw read data. After removing reads containing adapter sequences, poly A/T or G/C tracts, and low-quality reads, a total of 8.9 Gb of clean read data remained for *de novo* assembly. Using Trinity, all the sequencing reads were first assembled into 137 629 transcripts (11.12 Gb) with an average length of 808 bp, and N50 and N90 values of 1 424 and 305, respectively (Table 2). The largest clustered sequences obtained after Corset hierarchical clustering were assembled into unigenes, yielding 83 811 unigene sequences (9.5 Gb) with a mean length of 1 142 bp (Table 2). The N50 and N90 values for the unigenes were 1 676 bp and 529 bp, respectively; the lengths ranged from 201 to 13 577 bp, which was the same as that of the transcript data set (Table 2).

2.2 Evolutionary analysis of the CDS

A total of 79 535 CDS extracted from the unigenes provided information for phylogenetic analysis. The phylogenetic tree analysis was conducted based on the predicted protein sequences of 170 single-copy genes shared in *P. thomsonii*, alfalfa (*Medicago sativa*), and peanut *et al*.. For comparison, *P. lobata* and *M. sativa* clustered together in a single clade, suggesting that this species is phylogenetically closest to *M. sativa* (both are in the botanical family Fabaceae; Fig. 1).

2.3 Frequency distribution of the different types of SSR loci

The 83 811 unigenes were scanned by MISA, which identified a total of 25 452 SSR motifs in 21 052 unigene contigs. Among these, 5 023 contained tri-nucleotide repeats and were the most abundant, followed by 4 735 di-nucleotide repeats (Fig. 2).

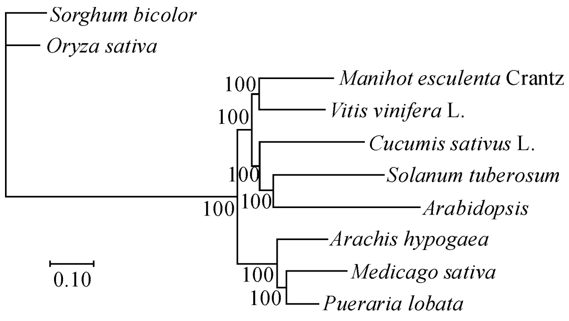


Fig. 1 Analysis of phylogeny of *P. lobata* (*P. thomsonii* Benth. 'No. 1') and other species

Table 2 *P. thomsonii* transcriptome assembly parameters

Item	Minimum length/bp	Maximum length/bp	Mean length/bp	N50	N90	Total nucleotides/bp
Transcripts	201	13 577	808	1 424	305	111 168 494
Unigenes	201	13 577	1 142	1 676	529	95 710 590

Note: N50/N90 indicate that the length of the assembly transcript is added to the length of the spliced transcript when it is no less than 50%/90% of total length

The tetra-, penta-, and hexa-nucleotide SSRs were 309, 64, and 21, respectively (Fig. 2). Also, SSR loci with fewer than five repeats were not expected to be included in the investigation. Among all the tri-nucleotide repeat units, (AAG)$_n$, (AAC)$_n$, and (AAT)$_n$ were the most common types with frequencies of 27.87% (1 400), 15.93% (800), and 16.44% (826), respectively (Table 3). Of all the AAG/CTT motifs with five repeat units, 726 (51.86%), were of the most abundant type. The AG/CT di-nucleotide repeat motif was the most abundant, and accounted for 71.68% (3 394) of all di-nucleotide repeats. The AC/GT repeat motif accounted for 20.76% (983) of the di-nucleotide repeats. Moreover, the AG/CT motif with six repeats represented the largest number of all the di-nucleotide repeats (Table 3), with a frequency of 33.71% (1 144).

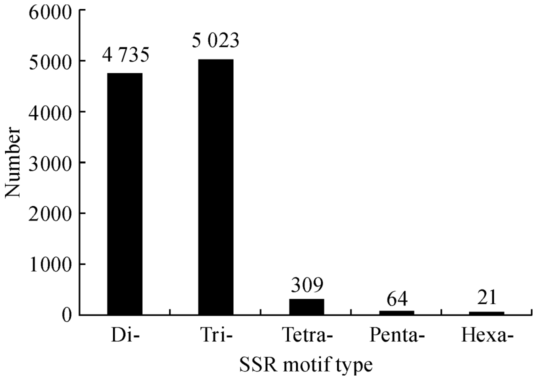


Fig. 2　The frequency of the SSR motifs in the *P. thomsonii* Benth. 'No. 1' transcriptome

2.4　SSR primer design and validation of SSR marker

A total of 16 574 pairs of primers were designed by Primer3 from the unique sequences flanking 25 452 SSR loci, with three pairs being designed for each SSR-containing sequence. The SSR motif loci longer than 18 bp and with PCR product sizes between 80 and 200 bp were selected for primer synthesis. Ultimately, one pair of primers for each SSR-containing sequence, 229 pairs of SSR primers in total, were synthesized for the genetic polymorphism test.

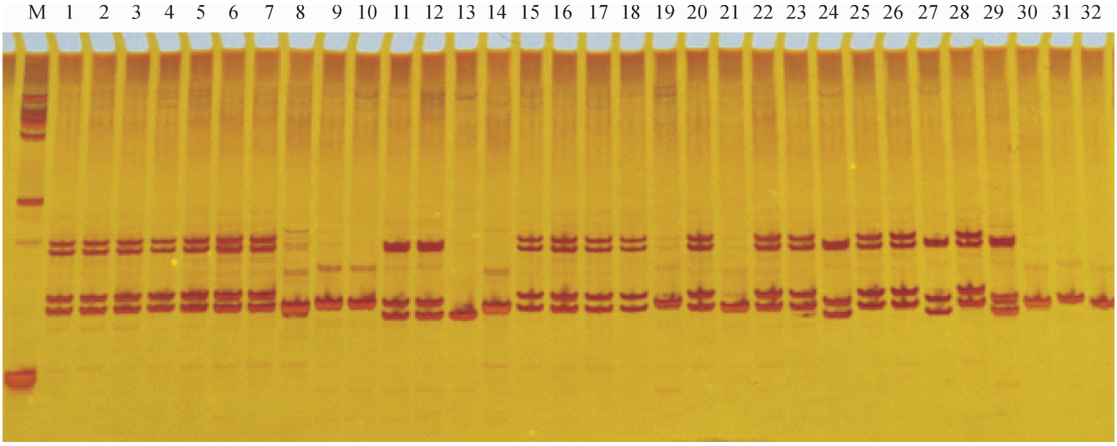
Initially, the 229 SSRs were scored for amplicon size polymorphism in six *Pueraria* collections. We found that 28 SSRs gave highly polymorphic and bands after PCR amplification. Fig 3 is shown the allelic variation of PtSSR144 in some *P. lobata* resources. These markers were used to determine the polymorphic information content (PIC) values and to estimate the genetic similarity of the 44 *Pueraria* germplasm collections (Table 4). A total of

Table 3　Frequency distribution of the number of di- and tri-nucleotide motif repeat units

Repeat	Number of motif repeats									Total
	5	6	7	8	9	10	11	12	12+	
AG/CT	—	1 144	817	584	423	291	129	5	1	3 394
AT/AT	—	390	199	193	157	131	54	2	—	1 126
AC/GT	—	398	255	140	107	40	29	14	—	983
CG/CG	—	6	—	2	—	—	—	—	—	8
AAG/CTT	726	461	206	6	1	—	—	—	—	1 400
AAT/ATT	408	261	150	5	1	—	—	—	1	826
AAC/GTT	443	239	105	11	—	2	—	—	—	800
ACC/GGT	371	153	88	7	—	—	—	—	—	619
ATC/ATG	346	189	75	2	—	—	—	—	—	612
AGG/CCT	270	128	51	6	—	—	—	—	—	455
AGC/CTG	254	96	47	6	—	—	—	—	—	403
CCG/CGG	234	81	29	1	—	—	—	—	—	345
ACG/CGT	75	55	29	1	—	—	—	—	1	161
ACT/AGT	67	27	13	1	—	—	—	—	—	108
Total	3 194	3 628	2 064	965	689	464	212	21	3	11 240

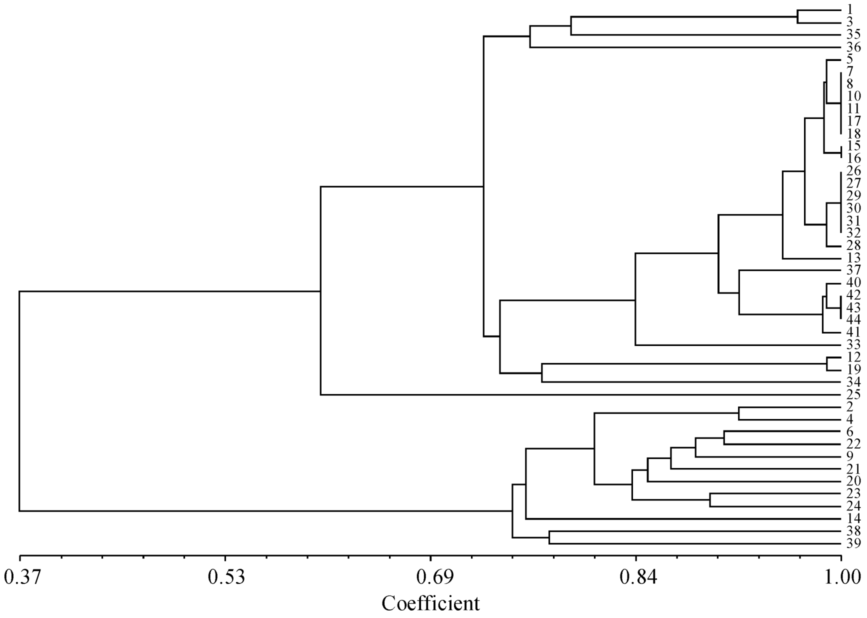
90 bands were amplified using the 28 SSRs in the 44 accessions，consisting of 89 alleles which gave a polymorphism rate 98.9%. The average number of alleles per SSR locus was 3.178 6. The PIC of the 28 markers ranged from 0.083 0 to 0.774 2 with average of 0.455 7. When PIC > 50%，the marker was highly polymorphic；when 25% < PIC < 50%，the marker was moderately polymorphic；when PIC < 25%，the marker had a low level of polymorphism[19]. In this study，10 SSRs，such as PtSSR108 and PtSSR122，were highly polymorphic，only PtSSR222 showed little polymorphism，and the remaining 17 markers were moderately polymorphic (Table 4).

The in-group similarity coefficient (0.58) was observed in the 44 germplasm accessions，and all accessions could be clearly divided into two groups；the first group included 32 accessions，and the second group included 12 (Fig. 4). Further in-group analysis revealed that the first group of 32 accessions contains five subclusters at a genetic similarity of 0.77；the second group of 12 accessions was divided into three subgroups，of them；accessions 36 and 25 in group 1 defined two subclusters that contained only a single individual. In group 2，accession number 14 was also in a unique



Line 1—32 indicated the 32 *P. lobata* resources；M indicated the DL2000 marker

Fig. 3 The marker PtSSR144 detecting 32 *P. lobata*



The number is as the same as in Table 1

Fig. 4 Cluster analysis of the 44 *Pueraria* accessions calculated from the 28 polymorphic genic-SSR marker alleles

Table 4　Analysis of polymorphic information content (PIC) and allele numbers in 28 *Pueraria* genic-SSR markers

SSR ID	Allele No.	Genotype No.	Major Allele Frequency	PIC
PtSSR36	2.000 0	3.000 0	0.772 7	0.289 6
PtSSR59	2.000 0	3.000 0	0.602 3	0.364 3
PtSSR98	2.000 0	3.000 0	0.681 8	0.365 6
PtSSR99	3.000 0	6.000 0	0.818 2	0.291 2
PtSSR104	6.000 0	7.000 0	0.738 6	0.410 2
PtSSR108	8.000 0	7.000 0	0.284 1	0.774 2
PtSSR109	3.000 0	4.000 0	0.613 6	0.468 5
PtSSR121	3.000 0	4.000 0	0.363 6	0.586 8
PtSSR122	8.000 0	8.000 0	0.295 5	0.758 4
PtSSR130	3.000 0	4.000 0	0.409 1	0.583 2
PtSSR135	5.000 0	4.000 0	0.340 9	0.629 5
PtSSR144	4.000 0	5.000 0	0.465 9	0.610 4
PtSSR155	2.000 0	3.000 0	0.602 3	0.364 3
PtSSR165	4.000 0	6.000 0	0.420 5	0.622 4
PtSSR168	3.000 0	5.000 0	0.750 0	0.326 6
PtSSR169	3.000 0	4.000 0	0.397 7	0.574 2
PtSSR172	2.000 0	3.000 0	0.613 6	0.361 8
PtSSR174	3.000 0	4.000 0	0.704 5	0.365 0
PtSSR175	3.000 0	4.000 0	0.636 4	0.427 8
PtSSR186	3.000 0	5.000 0	0.590 9	0.504 8
PtSSR187	5.000 0	6.000 0	0.500 0	0.593 3
PtSSR190	2.000 0	3.000 0	0.693 2	0.334 9
PtSSR191	4.000 0	4.000 0	0.738 6	0.354 1
PtSSR196	3.000 0	5.000 0	0.522 7	0.470 3
PtSSR201	3.000 0	5.000 0	0.625 0	0.441 5
PtSSR205	3.000 0	4.000 0	0.602 3	0.430 4
PtSSR217	2.000 0	3.000 0	0.534 1	0.373 8
PtSSR222	2.000 0	2.000 0	0.954 5	0.083 0
Mean	3.178 6	4.428 6	0.581 2	0.455 7

cluster. Based on the genetic similarity coefficient of 1.00, some accessions, such as numbers 7, 8, 10, 11, 17 and 18, were clustered together. Another six accessions (numbers 26, 27, 29, 30, 31, and 32) were also clustered together, indicating that there are no differences between them in this analysis. The same was for accessions 42, 43, and 44 (Fig. 4). It is indicated that the 44 germplasm resource did not show clear correlation to their geographic origin.

3　Discussion

Here, we investigated genetic diversity in a panel of 44 *Pueraria* accessions collected from the Guangxi area using SSR markers developed from RNA-seq data. SSR markers derived from a transcriptome database have some advantages over genomic SSRs, such as low cost and rapid development time. In addition, transcriptome sequencing data provides abundant gene function information, and genic SSRs are linked to the transcribed regions of the genome. For example, tri-nucleotide repeat units are associated with Huntington's disease in humans[20]. $(CT)_n$ repeats are closely linked to the *waxy* gene in rice, which encodes a granule-bound starch synthase and is correlated with grain amylose content[21]. Also, since genic-SSR markers are associated with transcribed regions of the genome, they are ideally suited for use in marker-assisted breeding.

In this study, tri-nucleotide SSR loci were found to be the most frequent (19.74%) followed by dinucleotide repeats (17.19%). This finding is similar to previous reports in cotton, pummelo, eggplant, and *Zanthoxylum bungeanum*[22-25], but contrasts with the distribution of genic-SSRs in spruce, pigeonpea, and *Gardenia jasminoides*[26-28]. $(AAG)_n$ were the most common type of tri-nucleotide repeat units, and the AG/CT motif was the most abundant of the di-nucleotide repeat motifs, similar to result found in pepper and coffee[29-30].

Jing *et al*[5] and Zhou *et al*[8] used RAPD markers to evaluate the genetic diversity of twelve *Pueraria* collections from the Chongqing area and eight from Jiangxi and Hunan provinces, and found polymorphism ratios for the two groups of 65.65% and 64.41%, respectively[5]. Chen *et al.*[6] analyzed the genetic relationships within 18 accessions of *P. thomsonii* Benth. using 22 pairs of SRAP markers, and found an average polymorphism ratio of 63.9%. The genetic distances ranged from 0.004 7 to 0.265 8. Eleven *Pueraria* accessions were genotyped with 12 polymorphic IS-

SR primers[7]. The genetic relationships within a group of 127 *Pueraria* accessions were analyzed with ISSR markers[9]. In this study, we found the polymorphism rate to be 98.9% for our SSR markers, which is much higher than that found for RAPD, ISSR, and SRAP markers in the studies mentioned above. This suggests that SSR marker is a powerful tool for the analysis of genetic diversity in *P. lobata* germplasm.

Pueraria accessions are mainly produced in the Guangxi area, especially in Tengxian and Wuzhou, and the varieties and quantity of germplasm resources are abundant in Guangxi. Our results show that there is no distinct boundary between *P. lobata* and *P. thomsonii*, which is consistent with the previous report[9]. For instance, accession number 40 shares a close relationship with numbers 42, 43, and 44, and both numbers 10 and 18, classified as *P. thomsonii*, have the same genetic background as four *P. lobata* accessions (7, 8, 11, and 17).

The genetic relationships between the 44 *Pueraria* accessions used in this study showed some correlations with geographic origin, although they were not absolute. For example, accession numbers 1 and 3 from Hechi, 7 and 8 from Baise, and 30, 31, and 32 from Huixian clustered together, although number 39 from Hechi was not included in this cluster. The *Pueraria* collections from Tengxian, numbers 12, 13, 14, 18, and 19, did not form a single cluster; this may be due to introductions from different regions, resulting in a certain degree of inconsistency between actual germplasm sources and clustering results[19]. In addition, we observed that the genetic similarities between most of the *Pueraria* accessions is greater than 0.73 in this study (Fig. 4), which suggests a narrow genetic base for these *Pueraria* accessions. Thus, it is urgent and necessary to increase the collection of *Pueraria* germplasm resources from other provinces in China to introduce diversity and broaden the genetic background of cultivated *Pueraria*.

In conclusion, using RNA-seq, we developed and validated 28 highly polymorphic SSR markers in a panel of 44 *Pueraria* accessions. Our work provides an informative and reproducible set of molecular markers for the evaluation of genetic relationships in kudzu, which will be useful for the establishment of a core germplasm collection of *Pueraria* species in the future.

Acknowledgements: The work was supported by the Guangxi Science and Technology Plan Project (grant number: Guike AD17195072) and Guangxi Science and Technology Major Project (grant number: Guike AA17202027-4) and Youth Fund of Guangxi Natural Science Foundation (grant number: 2018GXNSFBA294001).

References:

[1] HIEN T T, KIM H G, HAN E H, *et al*. Molecular mechanism of suppression of MDR1 by puerarin from *Pueraria lobata* via NF-kappaB pathway and cAMP-responsive element transcriptional activity-dependent up-regulation of AMP-activated protein kinase in breast cancer MCF-7/adr cells [J]. *Molecular Nutrition & Food Research*, 2010, **54**(7): 918-928.

[2] LIU B, WU Z Y, LI Y P, *et al*. Puerarin prevents cardiac hypertrophy induced by pressure overload through activation of autophagy[J]. *Biochemical and Biophysical Research Communications*, 2015, **464**(3): 908-915.

[3] YEUNG D K, LEUNG S W, XU Y C, *et al*. Puerarin, an isoflavonoid derived from *Radix* puerariae, potentiates endothelium-independent relaxation via the cyclic AMP pathway in porcine coronary artery[J]. *European Journal of Pharmacology*, 2006, **552**(1-3): 105-111.

[4] WANG J B. ISSR markers and their applications in plant genetics[J]. *Hereditas*, 2002, **24**(5): 613-616.

[5] JING X, XU L, CHEN J Y, *et al*. Genetic diversity of arrowroot (*Pueraria* L.) varieties revealed by RAPD analysis in Chongqing area [J]. *Chinese Agricultural Science Bulletin*, 2010, **26**(24): 80-82.

[6] CHEN D X, PENG R, LI L Y, *et al*. Analysis of genetic relationships of *Pueraria thomsonii* based on SRAP markers[J]. *China Journal of Chinese Material Medica*, 2011, **36**(5): 538-541.

[7] GUO Y Y, CHEN C Y, HUANG J L, *et al*. ISSR analysis of genetic relationships in *Radix* Puerariae from different original place [J]. *Popular Science & Technology*, 2013,**15**(4): 134-136.

[8] ZHOU J H, JIE Y C, DU X H, *et al*. RAPD analysis of arrowroot (*Pueraria* L.) germplasm genetic relationship [J]. *Crop Research*, 2013,**27**(4): 347-350.

[9] YUAN C, ZHONG WJ, GONG YY, *et al*. Genetic diversity and trait association analysis of *Pueraria lobata* resources [J]. *Journal of Plant Genetic Resources*, 2017,**18**(2): 233-241.

[10] POWELL W, MACHRAY G C, PROVAN J. Polymorphism revealed by simple sequence repeats[J]. *Trends in Plant Science*, 1996,**1**(7): 215-222.

[11] ZHANG L W, WAN X B, XU J T, *et al*. *De novo* assembly of kenaf (*Hibiscus cannabinus*) transcriptome using Illumina sequencing for gene discovery and marker identification [J]. *Molecular Breeding*, 2015,**35**(10): 192.

[12] GRABHERR M G, HAAS B J, YASSOUR M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome[J]. *Nature Biotechnology*, 2011,**29**(7): 644-652.

[13] DAVIDSON N M, OSHLACK A. Corset: enabling differential gene expression analysis for *de novo* assembled transcriptomes[J]. *Genome Biology*, 2014,**15**(7): 410.

[14] EDGAR R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput [J]. *Nucleic Acids Research*, 2004,**32**(5): 1 792-1 797.

[15] LI L, STOECKERT C J Jr, ROOS D S. OrthoMCL: identification of ortholog groups for eukaryotic genomes[J]. *Genome Research*, 2003,**13**(9): 2 178-2 189.

[16] KUMAR S, STECHER G, TAMURA K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets[J]. *Molecular Biology and Evolution*, 2016,**33**(7): 1 870-1 874.

[17] MURRAY M G, THOMPSON W F. Rapid isolation of high molecular weight plant DNA[J]. *Nucleic Acids Research*, 1980,**8**(19): 4 321-4 325.

[18] ZHANG L W, LI A Q, WANG X F, *et al*. Genetic diversity of kenaf (*Hibiscus cannabinus*) evaluated by inter-simple sequence repeat (ISSR)[J]. *Biochemical Genetics*, 2013,**51**(9-10): 800-810.

[19] NEI M. Molecular Evolutionary Genetics [M]. New York: Columbia University Press, 1987: 145-163.

[20] MACDONALD M. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes[J]. *Cell*, 1993,**72**(6): 971-983.

[21] AYERS N M, MCCLUNG A M, LARKIN P D, *et al*. Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm [J]. *Theoretical & Applied Genetics*, 1997,**94**(6-7): 773-781.

[22] ZHANG X W, YE Z W, WANG T K, *et al*. Characterization of the global transcriptome for cotton (*Gossypium hirsutum* L.) anther and development of SSR marker [J]. *Gene*, 2014,**551**(2): 206-213.

[23] LIANG M, YANG X M, LI H, *et al*. *De novo* transcriptome assembly of pummelo and molecular marker development[J]. *PLoS One*, 2015,**10**(3): e0120615.

[24] WEI M M, CHEN Y H, LIU F Z, *et al*. Development of SSR Markers for eggplant with transcriptome sequencing data [J]. *Journal of Plant Genetic Resources*, 2016,**7**(6): 1 082-1 091.

[25] FENG S J, ZHAO L L, LIU Z S, *et al*. *De novo* transcriptome assembly of *Zanthoxylum bungeanum* using Illumina sequencing for evolutionary analysis and simple sequence repeat marker development[J]. *Scientific Reports*, 2017,**7**(1): 16 754.

[26] RUNGIS D, BÉRUBÉ Y, ZHANG J, *et al*. Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags[J]. *Theoretical and Applied Genetics*, 2004,**109**(6): 1 283-1 294.

[27] DUTTA S, KUMAWAT G, SINGH B P, *et al*. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanu scajan* (L.) Millspaugh][J]. *BMC Plant Biology*, 2011,**11**(1): 17.

[28] DENG S Y, WANG X R, ZHU P L, *et al*. Development of polymorphic microsatellite markers in the medicinal plant *Gardenia jasminoides* (Rubiaceae)[J]. *Biochemical Systematics and Ecology*, 2015,**58**: 149-155.

[29] LIU F, WANG Y S, TIAN X L, *et al*. SSR mining in pepper (*Capsicum annuum* L.) transcriptome and the polymorphism analysis [J]. *Acta Horticulturae Sinica*, 2012,**39**(1): 168-174.

[30] AGGARWAL R K, HENDRE P S, VARSHNEY R K, *et al*. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species [J]. *Theoretical and Applied Genetics*, 2007,**114**(2): 359-372.

（编辑：宋亚珍）