Development and Application of simple sequence repeat markers based on whole-genome sequencing in Codonopsis lanceolata

Dan Zhang
Shanxi University of Chinese Medicine
Hongling Tian
Shanxi Agricultural University
Xiaoli Liu
Shanxi University of Chinese Medicine
Yuerong Zhang
Shanxi University of Chinese Medicine
Hui Li
Shanxi University of Chinese Medicine
Hai-xian Zhan (zhan030006@126.com)
Shanxi University of Chinese Medicine https://orcid.org/0000-0001-8204-5015

Research Article

Keywords: Codonopsis lanceolata, SSR marker, genetic diversity, clustering analysis, fingerprint code

Posted Date: May 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1513076/v1

Development and Application of simple sequence repeat markers based on whole-genome sequencing in *Codonopsis lanceolata*

Dan Zhang[1], Hongling Tian[2], Xiaoli Liu[1], Yuerong Zhang[1], Hui, Li[1], Haixian Zhan[1]

Abstract: The content of medicinal components of *Codonopsis pilosula* differ greatly in various habitats. The development of specific molecular markers is of great significance for the identification of different varieties of *C. pilosula*. In this study, 1,379,041 primer pairs were designed based on the whole genome of *C. lanceolata*. Of these, 6 pairs of primers were able to amplify polymorphic bands. We performed genetic diversity analysis and constructed a cluster tree and DNA fingerprint of the three groups: mixed, wild and cultivated *C. pilosula* materials. The results were as follows: first, the mixed materials had an average observed number of alleles (*Na*) of 7.6667, an average effective number of alleles (*Ne*) of 4.6058, an average Shannon's information index (*I*) of 1.7128, an average expected homozygosty (*Exp Hom*) of 0.2118, an average expected heterozygosity (*Exp Het*) of 0.7882, an average expected heterozygosity of Nei's (*Ht*) of 0.7798, and an average gene flow (*Nm*) of 0.1281. Second, the cultivated accessions were as follows: *Na* = 4.6667, *Ne* = 2.5956, *I* = 1.0702, *Exp Hom* = 0.4234, *Exp Het* = 0.5766, *Ht* = 0.5755, and *Nm* = 0.5253. Finally, the accessions of the wild materials were as follows: *Na* = 5.333, *Ne* = 2.8769, *I* = 1.2771, *Exp Hom* = 0.3556, *Exp Het* = 0.6444, *Ht* = 0.6433, and *Nm* = 0.5314. In addition, the results of cluster analysis showed that mixed, wild and cultivated materials were clustered into 4, 4 and 3 subgroups, respectively. Furthermore, we performed DNA fingerprinting of 47 mixed materials. These results are valuable for the identification and genetic analysis of *C. pilosula* from different sources in Shanxi Province.

Keywords: Codonopsis lanceolata, SSR marker, genetic diversity, clustering analysis, fingerprint code

1. Introduction

Codonopsis is the dried root of *Codonopsis pilosula* (Franch.) *Nannf*., *Codonopsis pilosula* Nannf. var. *modesta* (Nannf.) *L.T. Shen* or *Codonopsis tangshen* Oliv. *Codonopsis wall.* belonging to *Campanulaceae Juss. Codonopsis* is an herbal perennial dicotyledon. It consists of 42 species in Central Asia, East Asia, and South Asia, among which there are 40 species from China [1]. *Codonopsis* is a traditional medicine for strengthening the spleen, tonifying the lung, nourishing blood and promoting fluid; thus, it has been used to treat deficiency of the spleen and lung qi, insufficient qi and blood, internal heat to quench thirst, etc.

At present, the identification of the medicinal components of traditional Chinese medicines is considered most important for the evaluation of the quality of traditional Chinese medicine. There are great differences in the medicinal components of Chinese herbal medicines in different regions. This is mainly due to the genetic diversity caused by the origin environment, and gene flow has changed the content of traditional Chinese medicines. Therefore, it was necessary to search for a method to accurately identify C. *pilosula* in different areas. With the development of molecular marker technology, an increasing number of molecular markers have been used in medicinal plant research. These markers have been applied in genetic identification and genetic background analysis of species, such as *Rosa chinensis Jacq* [2], *Valeriana officinalis* L. [3], *and tartary buckwheat (Fagopyrum tataricum)* [4]. Among various types of markers, SSR markers have the advantages of high polymorphism, large number and codominance and have been widely used in the identification of many species, such as *Citrus* [5], *Dendrobium* [6], *Ophiophagus hannah* and *Naja atra* [7], *Capsicum chinensis Jacq.* [8], *Camellia sinensis* [9], *Medicago truncatula* [10], *Ziziphus jujuba* [11], and

Angelica gigas Nakai [12], *wheat* [13].

The molecular markers of C. *pilosula* materials mainly include RAPD [14-17], AFLPs (amplified fragment length polymorphisms) [18], SSRs (simple sequence repeats) [19], ISSRs (intersimple sequence repeats) [17,20], EST-SSRs (expressed sequence tags) [21], SRAPs (sequence-related amplified polymorphisms) [20], and DNA barcode technology [22-26]. However, these markers are not capable of meeting the needs of identification of C. *pilosula* resources. In addition, the whole-genome sequence of *C. lanceolata* was published in early 2021. These sequences will contribute to the development of molecular markers of *C. pilosula.* Many primers for other species, including *Platycodon grandiflorus* species, *Glycyrrhiza uralensis Fisch* species, *Docynia delavayi (Franch.)* species, and *Valeriana officinalis L.* species, have been developed based on genome-wide and transcriptome sequencing. These markers were applied to germplasm identification, diversity analysis and gene flow research.

In this study, we designed many markers for *C. lanceolata* based on the whole genome. Of these, 6 pairs of polymorphic markers were analyzed for genetic diversity, cluster tree construction and DNA fingerprinting of three groups of materials, including wild, cultivated and mixed materials from Shanxi Province. These results will be helpful to accurately identify *C. pilosula* in different regions. These markers will provide more references for further molecular marker-assisted breeding.

2. Materials and Methods

2.1 Materials and DNA Extraction

The fresh leaves or roots of the cultivars and wild materials of *C. pilosula* were collected from various regions in Shanxi Province (Table 1). The materials information is as follows. Fresh leaves and roots were dried by silica gel. Then, the samples were returned to the laboratory, and all of the collected materials were kept in a deep freezer (-80℃) for further research. Genomic DNA was extracted utilizing the SDS method [27]. Moreover, according to the requirements of the experiment, the collected accessions were divided into 3 groups, including 47 cultivated or wild mixed population materials (47 M), each consisting of 1 individual; 32 cultivated population samples (32 C), each consisting of 10 individuals; and 15 wild accession population accessions (15 W), each including 20 individuals, to launch diversity research.

Table 1 47 materials information

Latin name	Sampling location	Quantity	Growth pattern(C/W)	Longitude(N)	Latitude(E)	Altitude(m)
Codonopsis pilosula	Dongzhang Village, Pingshun County, Changzhi City	25	Cultivation	113°42′10.63″	36°08′51.83″	1440
C. pilosula	Siyuan Village, Qinyuan County, Changzhi City	25	C	112°32′99.62″	36°47′25.17″	1002
C. pilosula	Nangou Village, Pingshun County, Changzhi City	25	C	113°34′22.47″	35°98′11.76″	1381
C. pilosula	Nanshan Village, Wucheng Town, Lishi County, Luliang City	25	Wild	111°44′78.58″	37°42′18.95″	1435

C. pilosula	Hujia Mountain, Panquangou, Jiaocheng County, Luliang City	25	W	111°27′96.56″	37°84′90.77″	1473
C. pilosula	Maishan Ridge, Licheng County, Changzhi City	25	C	113°31′12.45″	36°50′12.53″	868
C. pilosula	Shijiapo Village, Lingchuan County, Jincheng City	25	C	113°44′41.77″	35°79′60.22″	1355
C. pilosula	Xiawujing Village, Pingshun County, Changzhi City	25	C	113°41′92.53″	36°30′04.72″	968
C. pilosula	Menlou Village, Pingshun County, Changzhi City	25	W	113°57′59.68″	36°13′43.87″	1372
C. pilosula	Menlou Village, Pingshun County, Changzhi City	25	W	113°41′04.66″	35°98′94.25″	1551
C. pilosula	Qinshanling, Licheng County, Changzhi City	25	C	113°30′92.83″	36°50′33.25″	826
C. pilosula	Duanjiazhuang, Qinyuan County, Changzhi City	25	W	112°28′55.51″	36°45′46.36″	1154
C. pilosula	Guanzhuang Village, Zezhou County, Jincheng City	25	C	112°76′60.47″	35°57′29.28″	970
C. pilosula	Nan Village, Licheng County, Changzhi City	25	C	113°34′57.89″	36°51′90.13″	826
C. pilosula	Menlou Village, Pingshun County, Changzhi City	25	C	113°57′85.99″	36°13′41.06″	1398
C. pilosula	Xia Village, Licheng County, Changzhi City	25	C	113°41′62.67″	36°50′67.93″	799
C. pilosula	Yangbai Village, Yangbai Township, Wutai County, Xinzhou City	25	C	113°08′53.12″	38°72′32.68″	850
C. pilosula	Liujiagou Village, Pingshun County, Changzhi City	25	C	113°47′45.52″	36°03′50.93″	1591
C. pilosula	Shijiapo Village, Lingchuan County, Jincheng City	25	C	113°45′53.05″	35°79′70.79″	1329

C. pilosula	Dongzhang Village, Pingshun County, Changzhi City	25	C	113°42′10.84″	36°08′52.27″	1440
C. pilosula	Xiaodongyu Village, Pingshun County, Changzhi City	25	C	113°45′89.63″	36°19′93.72″	1224
C. pilosula	Muhuayuan Village, Qinyuan County, Changzhi City	25	C	112°32′85.88″	36°46′34.66″	1004
C. pilosula	Yeshagou, Jiaocheng County, Luliang City	25	W	111°60′72.55″	37°78′58.94″	2125
C. pilosula	Yunding Mountain, Jiaocheng County, Luliang City	25	W	111°37′01.32″	37°87′75.40″	2172
C. pilosula	Beiwudang Mountain, Jiaocheng County, Luliang City	25	W	111°34′25.67″	37°77′40.02″	1586
C. pilosula	Changtiaogou, Jiaocheng County, Luliang City	25	W	111°43′61.65″	37°92′07.99″	1641
C. pilosula	Dongyi Village, Chengjiachuan Township, Lucheng District, Changzhi City	25	C	113°27′27.24″	36°28′90.57″	985
C. pilosula	Wangjia Village, Yangbai Township, Wutai County, Xinzhou City	25	C	113°08′34.17″	38°71′05.89″	867
C. pilosula	Zheng Village, Pingshun County, Changzhi City	25	C	113°42′25.05″	36°09′11.24″	1462
C. pilosula	Wudang Mountain, Shuijiaocheng County, Luliang	25	W	111°37′72.77″	37°78′06.79″	2006
C. pilosula	Beipo Village, Pingshun County, Changzhi City	25	C	113°40′91.98″	36°04′65.64″	1497
C. pilosula	Hetou Village, Changzi County, Changzhi City	25	C	112°87′41.81″	36°09′87.07″	939
C. pilosula	South of Old Gaoshan Village, Datong City	25	C	112°53′21.01″	40°04′54.53″	1261
C. pilosula	Shijiapo Village, Lingchuan County,	25	C	113°45′11.75″	35°79′81.88″	1430

C. pilosula	Jincheng City Dongpodi Township, Jiaocheng County, Luliang City	25	C	111°78′67.26″	37°68′35.56″	1174
C. pilosula	Shijiapo Village, Lingchuan County, Jincheng City	25	C	113°44′78.86″	35°79′80.57″	1431
C. pilosula	Shizui Township, Wutai County, Xinzhou City	25	C	113°41′40.01″	38°51′45.89″	1240
C. pilosula	South of Huabaotan Village, Zuoyun County, Datong City	25	C	112°38′57.28″	40°03′53.05″	1417
C. pilosula	Donghongya Village, Zhangjiachang Township, Zuoyun County, Datong City	25	C	112°46′59.90″	40°02′04.67″	1287
C. pilosula	Nanshan Temple, Wutai County, Xinzhou City	25	W	113°58′23.91″	38°98′19.57″	1831
C. pilosula	Yangqianbao Village, Zuoyun County, Datong City	25	C	112°49′35.47″	39°59′55.85″	1386
C. pilosula	Shijiapo Village, Lingchuan County, Jincheng City	25	W	113°44′77.78″	35°80′93.80″	1444
C. pilosula	Nanyukou Village, Wenshui County, Luliang City	25	W	111°99′35.26″	37°43′19.89″	847
C. pilosula	Nanlingdi Village, Pingyao County, Jinzhong City	25	W	112°33′95.92″	36°94′06.34″	1705
C. pilosula	Liuquan Township, Lingchuan County, Jincheng City	25	W	113°46′44.43″	35°79′38.86″	1342
C. pilosula	Zhangjing Village, Chengguan Town, Pingshun City, Changzhi City	25	C	113°45′53.32″	36°23′64.95″	1201
C. pilosula	Houziping Village, Hekou Township, Lan County, Luliang City	25	W	111°39′24.87″	38°29′39.54″	1634

2.2 Analysis of SSRs

According to the published SSR whole gene sequences of *C. lanceolata* and utilizing MISA

software (httpsa://webblast.ipk-gatersleben.de/misa/), we searched repeat sites of the whole gene sequences of *C. lanceolata.* The criteria were that the nucleotide base pairs had lengths of 1 bp, 2 bp, 3 bp, 4 bp, 5 bp and 6 bp. Simultaneously, those repeats had lengths of more than 6, 5, 5, 5, 5 and 5, respectively, and complex SSRs were less than 100 bp.

2.3 Primer Design

Primer 3-2.4.0 software was used to abstract the 300 bp sequences flanking the SSR repeat site as a template sequence for the marker. The design norm of the primers was as follows: 30% ~ 70% GC content, 100~250 bp product, Tm value of 55 ℃ ~ 65 ℃, upstream and downstream primers with Tm value less than 5 ℃, and 18~22 bp primer length. In this study, the primers used were synthesized by Sangon Biotech Co., Ltd., China (www.sangon.com).

2.4 Validation of Primers

2.5 Genetic Diversity Analysis

The data matrix was constructed by the genotype statistical method. The diversity data in the exported cluster tree file were analyzed using popgen32 software. Furthermore, we utilized MEGA-X and R language to draw clustering tree. The values of genetic diversity analysis include genetic variation (observed number of alleles (*Na*), effective number of alleles (*Ne*) and Shannon's information index (*I*)), heterozygous value (expected homozygote, expected heterozygote, expected heterozygosity of Nei's), F statistics, and *Nm*.

2.6 Construction of DNA fingerprint code

The DNA fingerprint contained sample numbers, amplified sites and a data matrix of “0,1”. “0” indicates that there was no amplification site. In contrast, “1” indicates 1 amplification site. The special fingerprint code of each material can be used to identify the sample.

3 Results

3.1 Repeat Types and Frequency of SSRs

The whole-genome sequence of *C. lanceolata was* approximately 1.2 Gb. These sequences were anchored on 8 chromosomes and included various SSR sites distributed between 47271 and 86028 base sequences, and the average site was between 454.25 Mb^{-1} and 475.10 Mb^{-1}. In addition, SSR sites of 1-6 nucleotides were universal in all 8 chromosomes. The number of repeat units was negatively correlated with the distribution of SSR loci in the whole genome. Moreover, mononucleotides accounted for the majority of SSR loci, followed by dinucleotides and trinucleotides. In total, mononucleotides and dinucleotides accounted for a considerable proportion of the SSR loci. The suggested that *C. lanceolata* may be a species with late origin and high evolution

MISA and Primer 5.0 software were used to analyze repeat SSR loci and the designed primers. We designed 1379041 primer pairs for *C. lanceolata.* A total of 350 primer pairs distributed on 8 chromosomes were synthesized randomly, of which 270 primer pairs could amplify bands. Further analysis showed that six pairs of primers located on chromosomes 5 and 7 amplified polymorphic bands among different *C. pilosula* varieties (Fig.1).

Fig.1 Amplification results of 6 polymorphism primers

3.3 Analysis of Genetic Relationships

All gene variants of 47 mixed samples were analyzed by POPGENE software (Table 2). The results showed that the *Na* value ranged from 7 to 9 (mean: 7.6667), the *Ne* value ranged from 3.7251 to 5.1915 (mean: 4.6085), and the *I* value ranged from 1.6163 to 1.8584 (mean: 1.7128). Furthermore, the heterozygosity statistics of all loci revealed that the expected homozygosity range was 0.1839 to

0.2606 (mean: 0.2118), the expected heterozygosity range was 0.7394 to 0.8161 (mean: 0.7882), and the expected heterozygosity range of Nei was 0.7316 to 0.8074 (mean: 0.7798). At the same time, F-statistics and gene flow (*Nm*) of all loci demonstrated that Fst ranged from 0.5201 to 0.8146 (average: 0.6612) and *Nm* ranged from 0.0569 to 0.2307 (average: 0.1281). Ultimately, the genetic distance ranged from 0.0476 to 0.8168, and the genetic identity ranged from 0.4419 to 0.9535. According to the maximum genetic distance (0.8168) and the minimum genetic identity (0.4419), the 47 mixed samples were divided into five groups. Nangou Village, Pingshun County, Changzhi City population and Yunding Mountain, Jiaocheng County, Luliang City population, Nangou Village, Pingshun County, Changzhi City population and Nanyukou Village, Wenshui County, Luliang City population, Nangou Village, Pingshun County, Changzhi City population and Houziping Village, Hekou Township, Lan County, Luliang City population, Hetou Village, Changzi County, Changzhi City population and Nanlingdi Village, Pingyao County, Jinzhong City population, as well as Nanshan Temple, Wutai County, Xinzhou City population and Nanlingdi Village, Pingyao County, Jinzhong City population. Based on the minimum genetic distance (0.0476) and the maximum genetic identity (0.9535), the 47 mixed materials were divided into two groups. Dongpodi Township, Jiaocheng County, Luliang City population and Nanshan Temple, Wutai County, Xinzhou City population, Shizui Township, Wutai County, Xinzhou City population and Nanshan Temple, Wutai County, Xinzhou City population.

Then, the *Na* value ranged from 3 to 6 (average: 4.6667), the *Ne* value ranged from 1.6497 to 3.5369 (average: 2.5956), and the *I* value ranged from 0.5928 to 1.4527 (average: 1.0702) in the 32 cultured accessions. Next, a summary of heterozygosity statistics for all loci revealed that the expected homozygosity ranged from 0.2814 to 0.6054 (average: 0.4234), the expected heterozygosity ranged from 0.3946 to 0.7186 (average: 0.5766), and Nei's expected heterozygosity ranged from 0.3938 to 0.7173 (average: 0.5755). Moreover, the results of F-statistics and gene flow (*Nm*) for all loci demonstrated that the *Fst* value ranged from 0.1682 to 0.5180 (average: 0.3224), and the *Nm* value ranged from 0.2327 to 1.2366 (average: 0.5253). Nei's value of genetic identity and genetic distance indicated that the genetic distance ranged from 0.0457 to 0.4770, and the genetic identity ranged from 0.6206 to 0.9553.

According to the highest genetic distance (0.4770) and the lowest genetic identity (0.6206) of the 32 cultured accessions, we found that Guanzhuang Village and Zheng Village had pairwise materials. Similarly, the lowest genetic distance (0.0457) and the highest genetic identity (0.9553) demonstrated that there were pairwise materials in the Shizui Township and Datong City populations.

In addition, the accessions for the 15 wild materials were as follows: *Na* >5 and <6 (average: 5.333), *Ne* >2.4717 and <3.5584 (average: 2.8769), and *I* >1.1718 and <1.4876 (average: 1.2771). Next, heterozygosity statistics for all loci revealed that the expected homozygosity was >0.2798 and <0.4036 (average: 0.3556), the expected heterozygosity was >0.5964 and <0.7202 (average: 0.6444), and the Nei's expected heterozygosity was >0.5954 and <0.7190 (average: 0.6433). Third, F-statistics and gene flow (*Nm*) for all loci showed that the *Fst* was >0.1411 and <0.4931 (average: 0.3200) and the *Nm* was >0.2570 and <1.5221 (average: 0.5314).

Ultimately, Nei's original measures of genetic identity and genetic distance showed that the genetic distance was between 0.0353 and 0.7001 and that the genetic identity was between 0.4965 and 0.9654. On the basis of the highest genetic distance (0.7001) and the lowest genetic identity (0.4965), the materials of Hujia Mountain and Wenshui County populations were pairwise. Similarly, the lowest genetic distance (0.0353) and the highest genetic identity (0.9654) revealed that the materials from Luliang City and Luliang City were pairwise populations.

Table 2 Genetic diversity statistics

	Type	Na	Ne	I	Exp Hom	Exp Het	Ht	Fst	Nm
	M	7.6667	4.6085	1.7128	0.2118	0.7882	0.7798	0.6612	0.1281
Mean	32C	4.6667	2.5956	1.0702	0.4234	0.5766	0.5755	0.3224	0.5253
	15W	5.3333	2.8769	1.2771	0.3556	0.6444	0.6433	0.3200	0.5314

3.4 Cluster Analysis

The 47 mixed materials were divided into 4 subgroups by cluster analysis (Fig.2). Subgroup I samples included Changzhi City, Xinzhou City, Luliang City and Jincheng City samples. Subgroup II included Xinzhou City, Lvliang City, Datong City, Jincheng City, and Changzhi City samples. Likewise, subgroup III comprised Changzhi City, Jinzhong City, and Jincheng City samples. Subgroup IV consisted of Lvliang City and Jincheng City samples. The above results further illustrated that the sources of *C. pilosula* had minor diversity in 47 regions of Shanxi Province.

The 32 cultured materials were divided into 4 subgroups by cluster analysis. Subgroup I samples included Changzhi City, Xinzhou City, and Jincheng City samples. Subgroup II included Datong City, Xinzhou City, Lvliang City and Jincheng City samples. Likewise, subgroup III comprised Changzhi City and Jincheng City samples. Subgroup IV consisted of Changzhi City and Jincheng City samples.

The 15 wild materials were divided into 3 subgroups by cluster analysis. Subgroup I samples included Lvliang City, Changzhi City, and Jincheng City samples. Subgroup II included Lvliang City and Jinzhong City. samples. Likewise, Subgroup III consisted of Jincheng City samples.

Fig.2 Circular clustering tree. **A** 47 mixed materials. **B** 32 cultivated materials.**C** 15 wild materials

3.5 DNA Fingerprint Analysis

The materials in 47 different regions had unique DNA fingerprints, so each material was accurately identified (Fig.3). Further analysis showed that the similarity of DNA fingerprints of individual materials was high; for example, the similarity of materials 8 and 10 was high, but the two materials were distinguishable by examining the specific bands amplified by three pairs of primers, dsq-5, dsq5-36 and dsq7-38. Although the fingerprints of J-3 and J-4 also had high similarity, the two materials were distinguishable by six primers. In summary, the DNA fingerprints constructed in this study could be used for the accurate identification of different *C. pilosula* sources in Shanxi Province.

Fig.3 DNA fingerprints of 47 mixed materials

4. Discussion

4.1 Genetic Diversity

Genetic diversity is considered to be an important part of biodiversity and the essence of species evolution. If a population or species has higher gene diversity or abundant gene variation [28], they have a strong ability to adapt to the environment. Evaluation of genetic diversity promotes the maintenance of diverse species/cultivars for preservation and crop amelioration [29,30]. Thirty-nine SSR markers were used to analyze the sixteen *C. lanceolata* accessions from Korea. The results showed that the average *Na* was 5.08, the average *He* was 0.59 and the average *Ho* was 0.48 [31]. The genetic diversity of *Cannabis* species was calculated by 49 pairs of primers; the average *Na* was 4.13, the average *Ne* was 2.50, the average *I* was 1.01 and the average *He* was 0.56 in 12 *Cannabis* varieties [32]. Twenty-five polymorphic SSR markers were utilized in 18 natural populations (425 individuals) of *Tectona grandis* L.f. in Indian species, and these markers showed that the average *Na* was 7.70, the average *Ne* was 4.30, the average *I* was 1.49, and the average *Ho* and He were 0.614 and 0.668,

respectively [33]. Thirty-six collected SSR markers were amplified from 48 various *spinach* accessions for molecular validation, and the results showed that the average *Na* was 2.9 and the average Nei's expected heterozygosity was 0.4 [34]. Seventy-three polymorphic primers accurately identified 37 Dendrobium species and demonstrated that the average *He* was 0.375, and the average *Nm* was 0.753 [35]. In this study, the whole summary of genetic variation statistics for all loci decreased in the order M > W > C, the whole summary of heterozygosity statistics for all loci decreased in the order M > W > C, and the whole summary of F-statistics and gene flow for all loci decreased in the order W > C > M. Combined with the above analysis, each locus possessed abundant levels of polymorphism, showing that wild samples frequently had gene interactions; in contrast, mixed materials showed minimum gene communication.

4.2 Cluster Research

When some individuals migrate from one group to another, they bring their genes to the new group, resulting in gene flow. Gene flow is generally considered to be an important factor in the degree of genetic interspecific and intraspecific variation. Therefore, the greater the gene flow, the greater the similarity between the two populations [36]. Wright proposed that if the gene flow value between populations is less than 1 (Nm<1), limited gene flow is the main reason for genetic differentiation [37,38]. In this study, the average *Nm* values of cultivated and wild materials were 0.5253 and 0.5314, respectively, which showed that there was great genetic similarity between the two groups. Most importantly, gene flow is affected by many factors, such as the direction and speed of wind and water, individual survival time, climate and geographical obstacles, which affect the dispersion of the population. Some researchers believe that habitat fragmentation affects gene flow [39-41]. In our study, wild materials of *C.* lanceolata mainly relies on bees, wind and rain to spread pollen, and after being digested as food for animals, the fruit takes root and sprouts in the form of feces. Cultivated accessions, after unified collection through human intervention, may be planted and managed in the same field or greenhouse. In addition, fragmentation of habitat can also affect gene flow, and this is caused by human activities and gradually develops with increasing intensity of human activity. The gene flow was relatively small (0.1281) in the mixed materials, which may be due to geographical obstacles or habitat fragmentation. The results of the experiment will further provide references for the identification of germplasm resources in *C. pilosula*.

4.3 Fingerprint Code

DNA fingerprints had multiple loci, high variability, and somatic cell stability. They play important roles in variety identification and breeding [42]. Furthermore, DNA fingerprint databases are essential and efficient measures for molecular research of plants that provide accurate information for breeding, variety quality analysis, variety resource protection, and molecular marker-assisted breeding [43,44]. SSRs have been successfully used to construct DNA fingerprints and can be easily and efficiently used for the identification of multiple other species [45], such as Asiatic cotton varieties of India [46], chili pepper varieties [47], leaves of Datura stramonium [48], strawberry varieties [49], and aromatic rice [50]. Therefore, the DNA fingerprint code developed for *C. pilosula* in this study can be used as the basis for species resource identification. At the same time, the results of this study could also serve as a reference for intuitively revealing the genetic diversity and genetic relationships among different sources of *C. pilosula*. Meanwhile, the unique DNA fingerprints of 47 mixed materials were constructed using 6 pairs of polymorphic primers to accurately identify *C. pilosula* from different areas. The above results can establish a method for the construction and improvement of a DNA fingerprint database of *C. pilosula* variety sources.

Conclusion

C. *pilosula* is one of the medicine and food homology directories. Based on the whole genome information of *C. lanceolata*, we analyzed genetic diversity and constructed a cluster tree and DNA fingerprint of collected accessions by designing six pairs of polymorphic primers. This study will contribute to research on *C. pilosula* in wild and cultivated varieties, molecular marker-assisted breeding, and genetic diversity.

Supplementary Information

The online version contains supplementary material available at document 1-3.

Author contributions

HZ contributed to the theme of the study; HZ, DZ, HT, YZ and HL collected samples; DZ performed the experiment; HZ, DZ, XL and HL performed the data analysis. DZ drafted the manuscript. HZ edited the manuscript. All coauthors read and approved the final manuscript.

Declarations

Conflict of interest

The authors declare that they have no competing interests.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors

Funding

This work was supported by Shanxi Provincial Program of key R & D projects (China) (No. 201903D221053), the earmarked fund for Modern Agro-industry Technology Research System (China) (No. 2022) and Doctoral foundation of Shanxi University of Chinese Medicine (China) (2020BK05).

Reference

1. He JY, Ma N, Zhu S, Komatsu K, Li ZY, Fu WM (2015) The genus codonopsis (campanulaceae): a review of phytochemistry, bioactivity and quality control. Journal of Natural Medicines, 69(1):1-21

2. Xia AN, Yang AA, Meng XS, Dong GZ, Tang XJ, Lei SM, Liu YG (2022) Development and application of rose (rosa chinensis jacq.) snp markers based on SLAF-seq technology. Genetic Resources and Crop Evolution, 69(1):73-182

3. Boczkowska M, Bczek K, Kosakowska O, Anna R, Wglarz Z (2020) Genome-wide diversity analysis of valeriana officinalis l. using dart-seq derived snp markers. Agronomy, 10(9):1346

4. Shi TX, Li RY, Zheng R, Chen QF, Liang CG (2021) Mapping QTLs for 1000-grain weight and genes controlling hull type using snp marker in tartary buckwheat (fagopyrum tataricum). BMC Genomics, 22(1):142

5. Duhan N, Meshram M, Loaiza CD, Kaundal R (2020) Citsatdb: genome-wide simple sequence repeat (SSR) marker database of citrus species for germplasm characterization and crop improvement.

Genes, 11(12):1486

6. Zhao TM, Zheng SG, Hu YD, Zhao RX, Li HJ, Zhang XQ, Chun Z (2019) Classification of interspecific and intraspecific species by genome-wide SSR markers on dendrobium. South African Journal of Botany, 127:136-146

7. Liu WC, Xu YT, Li ZK, Fan J, Yang Y (2019) Genome-wide mining of microsatellites in king cobra (Ophiophagus hannah) and cross-species development of tetranucleotide SSR markers in Chinese cobra (Naja atra). Molecular biology reports,46(6):6087-6098

8. Uncu AT (2019) Genome-wide identification of simple sequence repeat (SSR) markers in capsicum chinense jacq. with high potential for use in pepper introgression breeding. Biologia, 74(2):119-126

9. Liu SR, An YL, Li FD, Li SJ, Liu LL, Zhou QY, Zhao SQ, Wei CL (2018) Genome-wide identification of simple sequence repeats and development of polymorphic SSR markers for genetic studies in tea plant (Camellia sinensis). Molecular Breeding, 38(5):1-13

10. Min XY, Zhang ZS, Liu YS, Wei XY, Liu ZP, Wang YR, Liu WX (2017) Genome-Wide Development of MicroRNA-Based SSR Markers in Medicago truncatula with Their Transferability Analysis and Utilization in Related Legume Species. International Journal of Molecular Sciences, 18(11):2440

11. Chen W, Hou L, Zhang ZY, Pang XM, Li YY (2017) Genetic diversity, population structure, and linkage disequilibrium of a core collection of ziziphus jujuba assessed with genome-wide snps developed by genotyping-by-sequencing and SSR markers. Frontiers in Plant Science, 8:575

12. Gil J, Um Y, Kim S, Kim OT, Koo SC, Reddy CS, Lee Y (2017) Development of genome-wide SSR markers from Angelica gigas Nakai using next generation sequencing. Genes, 8(10):238

13. Li L, Sun FY, Wu D, Zhen F, Bai GH, Gao DR, Li T (2017) High-throughput development of genome-wide locus-specific informative SSR markers in wheat. Science China (Life Sciences) (06):671-673

14. Hong BG, Bao R L, Qian HW, Jia KC, Tong SZ (2007) Abundant genetic diversity in cultivated codonopsis pilosula populations revealed by rapd polymorphisms. Genetic Resources & Crop Evolution, 54(5):917-924

15. Zhang JQ, Xue SU, Qiong WU, Ding SS, Sun K (2006) Analysis of rapd on medicinal plants of codonopsis pilosula. Zhong yao cai=Zhongyaocai=Journal of Chinese medicinal materials, 29(5):417-420

16. Park JH, Kim YH, Nou IS, Choi JE, Shim IY, Kang KK (1997) Evaluation of genetic diversity among Korean wild Codonopsis lanceolata by using RAPD. Korean J. Plant. Res, 10(3):258-264

17. Guo WL, Gong L, Ding ZF, Li YD, Li FX, Zhao SP, Liu B (2006) Genomic instability in phenotypically normal regenerants of medicinal plant Codonopsis lanceolata Benth. et Hook. f., as revealed by ISSR and RAPD markers. Plant cell reports, 25(9):896-906

18. Gu C, Cao LY, Su Q, Guan LJ, Yang J, Gao JP (2016) AFLP and HPLC Fingerprints A nalysis of Codonopsis Species from Original Areas and the Same Planting Base. Zhong yao cai =Zhongyaocai=Journal of Chinese Medicinal Materials, 39(8):1716-1722

19. Kim S, Ji HJ, Chung H, Ji HK, Yi L (2016) Simple sequence repeat marker development from codonopsis lanceolata and genetic relation analysis. Journal of Plant Biotechnology, 43(2):181-188

20. Chen D, Rui P, Li L, Sun N, Cai Y (2009) Study on genetic diversity of codonopsis tangshen by SRAP and ISSR markers. Zhongguo Zhong Yao Za Zhi, 34(3):255-259

21. Lee J, Gil J, Um Y, Kim HB, Kim SC, Koo S, Lee Y (2017) Genetic diversity analysis of codonopsis species using codonopsis laneolata based EST-SSR markers. 한국약용작물학회 학술대회논문집, 25(1):59-59

22. Wang DY, Wang Q, Wang YL, Xiang XG, Huang LQ, Jin XH (2017) Evaluation of DNA barcodes in Codonopsis (Campanulaceae) and in some large angiosperm plant genera. PloS one, 12(2): e0170286

23. He JY, Zhu S, Komatsu K, Yukihiro G (2014) Genetic polymorphism of medicinally-used codonopsis species in an internal transcribed spacer sequence of nuclear ribosomal dna and its application to authenticate codonopsis radix. J Nat Med, 68(1):112-124

24. Lin TC, Hsieh CC, Agrawal CR, Kuo CL, Chueh FS, Tsay HS (2007) ITS sequence based phylogenetic relationship of dangshen radix. J Food Drug Anal, 15(4):428-432

25. Hwang SG, Ju HK, Moon JC, Kim JH, Jang CS (2016) Comparative analysis of chloroplast dna sequences of codonopsis lanceolata and platycodon grandiflorus and application in development of molecular markers. Applied Biological Chemistry, 60(1):1-9

26. Zhao S, Xin T, Hou D, Pang X, Chen R, Gao J (2013) Identification of codonopsis Radix and its adulterants using the ITS/ITS2 barcodes. World Science and Technology-Modernization of Traditional Chinese Medicine, 421-428

27. Thompson BM, Binkley JM, Stewart GC (2011) Current physical and SDS extraction methods do

342 not efficiently remove exosporium proteins from Bacillus anthracis spores. Journal of microbiological
343 methods, 85(2):143-148
344 28. Holderegger R, Balkenhol N, Bolliger J, Engler JO, Gugerli F, Hochkirch A, Nowak C,
345 Segelbacher G, Widmer A, Zachos FE (2019) Conservation genetics: linking science with practice.
346 Molecular Ecology, 28(17).
347 29. Ppk A, Tj A, Kvb B (2020) Microsatellite based dna fingerprinting and assessment of genetic
348 diversity in bougainvillea cultivars - sciencedirect. Gene, 753:144794
349 30. Aremu CO (2011) Genetic diversity: a review for need and measurements for intraspecies crop
350 improvement. Journal of Microbiology and Biotechnology Research, 1(2):80-85
351 31. Kim S, Jo N, Gil JS, Koo SC, Um Y, Hong CP, Lee Y (2021) Development of genome-wide simple
352 sequence repeat markers in Codonopsis lanceolata using next-generation sequencing. Horticulture,
353 Environment, and Biotechnology, 62(6):985-993
354 32. Zhang L, Saha D, Zhang J, Zhao L, Zhang C (2020) Citation: genetic diversity and population
355 structure of cannabis based on the genome-wide development of simple sequence repeat markers.
356 Frontiers in Genetics, 11:958
357 33. Balakrishnan S, Dev SA, Sakthi AR, Vikashini B, Reshma BT, Magesh NS, Ramasamy Y (2021)
358 Gene-ecological zonation and population genetic structure of tectona grandis l.f. in india revealed by
359 genome-wide ssr markers. Tree Genetics & Genomes, 17(4):1-14
360 34. Bhattarai G, Shi A, Kandel DR, Solís-Gracia N, Da Silva JA, Avila CA (2021) Genome-wide
361 simple sequence repeats (SSR) markers discovered from whole-genome sequence comparisons of
362 multiple spinach accessions. Scientific reports, 11(1):1-16
363 35. Zhao TM, Zheng SG, Hu YD, Zhao RX, Li HJ, Zhang XQ, Chun Z (2019) Classification of
364 interspecific and intraspecific species by genome-wide SSR markers on dendrobium. South African
365 Journal of Botany, 127:136-146
366 36. Williams KR, Wasson SR, Barrett A, Greenall RF, Jones SR, Bailey EG (2021) Teaching
367 Hardy-Weinberg Equilibrium using Population-Level Punnett Squares: Facilitating Calculation for
368 Students with Math Anxiety. CBE life sciences education, 20(2): ar22
369 37. Kwong AM, Blackwell TW, LeFaive J, De Andrade M, Barnard J, Barnes KC, Kang HM (2021)
370 Robust, flexible, and scalable tests for Hardy–Weinberg equilibrium across diverse
371 ancestries. Genetics, 218(1): iyab044

372 38. Graffelman J, Ortoleva L (2020) A network algorithm for the X chromosomal exact test for
373 Hardy-Weinberg equilibrium with multiple alleles. Molecular ecology resources, 21(5):1547-1557
374 39. Kritika MG, Balaji C (2021) Gene Flow in Volant Vertebrates: Species Biology, Ecology and
375 Climate Change. Journal of the Indian Institute of Science(prepublish), 101(2):165-176
376 40. Cui D, Tang C, Lu H, Li J, Han L (2021) Genetic differentiation and restricted gene flow in rice
377 landraces from yunnan, china: effects of isolation-by-distance and isolation-by-environment. Rice,
378 14(1):1-14
379 41. González AV, Gómez-Silva V, Ramírez MJ, Fontúrbel FE (2020) Meta-analysis of the differential
380 effects of habitat fragmentation and degradation on plant genetic diversity. Conservation
381 Biology, 34(3):711-720
382 42. Backiyarani S, Chandrasekar A, Uma S, Saraswathi MS (2019) MusatransSSRDB (a transcriptome
383 derived SSR database)–An advanced tool for banana improvement. *Journal of biosciences*, *44*(1):1-10
384 43. Wilton R, Wheelan SJ, Szalay AS, Salzberg SL (2019) The Terabase Search Engine: A large-scale
385 relational database of short-read sequences. *Bioinformatics*, *35*(4):665-670
386 44. Jayashree B, Reddy PT, Leeladevi Y, Crouch JH, Mahalakshmi V, Buhariwalla HK, Hoisington DA
387 (2006) Laboratory Information Management Software for genotyping workflows: applications in high
388 throughput crop genotyping. *BMC bioinformatics*, *7*(1):1-6
389 45. Freixas-Coutin JA, An S, Postman J, Bassil NV, Yates B, Shukla M, Saxena PK (2019)
390 Development of a reliable Corylus sp. reference database through the implementation of a DNA
391 fingerprinting test. *Planta*, *249*(6):1863-1874
392 46. Santosh HB, Meshram M, Santhy V, Waghmare VN (2021) Microsatellite marker-based diversity
393 analysis and dna fingerprinting of asiatic cotton (Gossypium arboreum) varieties of india. Journal of
394 Plant Biochemistry and Biotechnology (3):1-8
395 47. Terryana RT, Rijzaani H, Priyatno TP, Manzila I, Lestari P (2020, March) Construction of DNA
396 fingerprint for chili pepper varieties using SNAP markers. IOP Conference Series: Earth and
397 Environmental Science, 482(1): 012038
398 48. Arshad H, Shadma W, Moustafa M (2020) Pharmacognostic standardization and dna fingerprinting
399 of leaves of datura stramonium, growing naturally in asir region of saudi arabia. Pakistan Journal of
400 Pharmaceutical Sciences, 33(3):1155-1161
401 49. Aguilar TM, Calderón ZG, Gutiérrez Espinosa MA, Lobato Ortíz R, Córdoba Téllez L, Volke

Haller V, Santacruz Varela A (2019) DNA fingerprint of strawberry varieties developed at Colegio de Postgraduados, Mexico. *Nova scientia*, *11*(22):186-206

50. Haque MS, Tabassum T, Saha NR, Islam MS (2019) Dna fingerprint and diversity in aromatic rice by gn1 gene linked ssr markers. Progressive Agriculture, 29(4):336-344

Authors and Affiliations

Dan Zhang[1], Hongling Tian[2], Xiaoli Liu[1], Yuerong Zhang[1], Hui Li[1], Haixian Zhan[1]

✉ Haixian Zhan

Zhan030006@126.com

[1] Institute of Pharmaceutical & Food engineering, Shanxi University of Chinese Medicine, Jingzhong 030619, China

[2] Institute of Economic Crops, Shanxi Agricultural University, Fenyang 032200, China

Figures

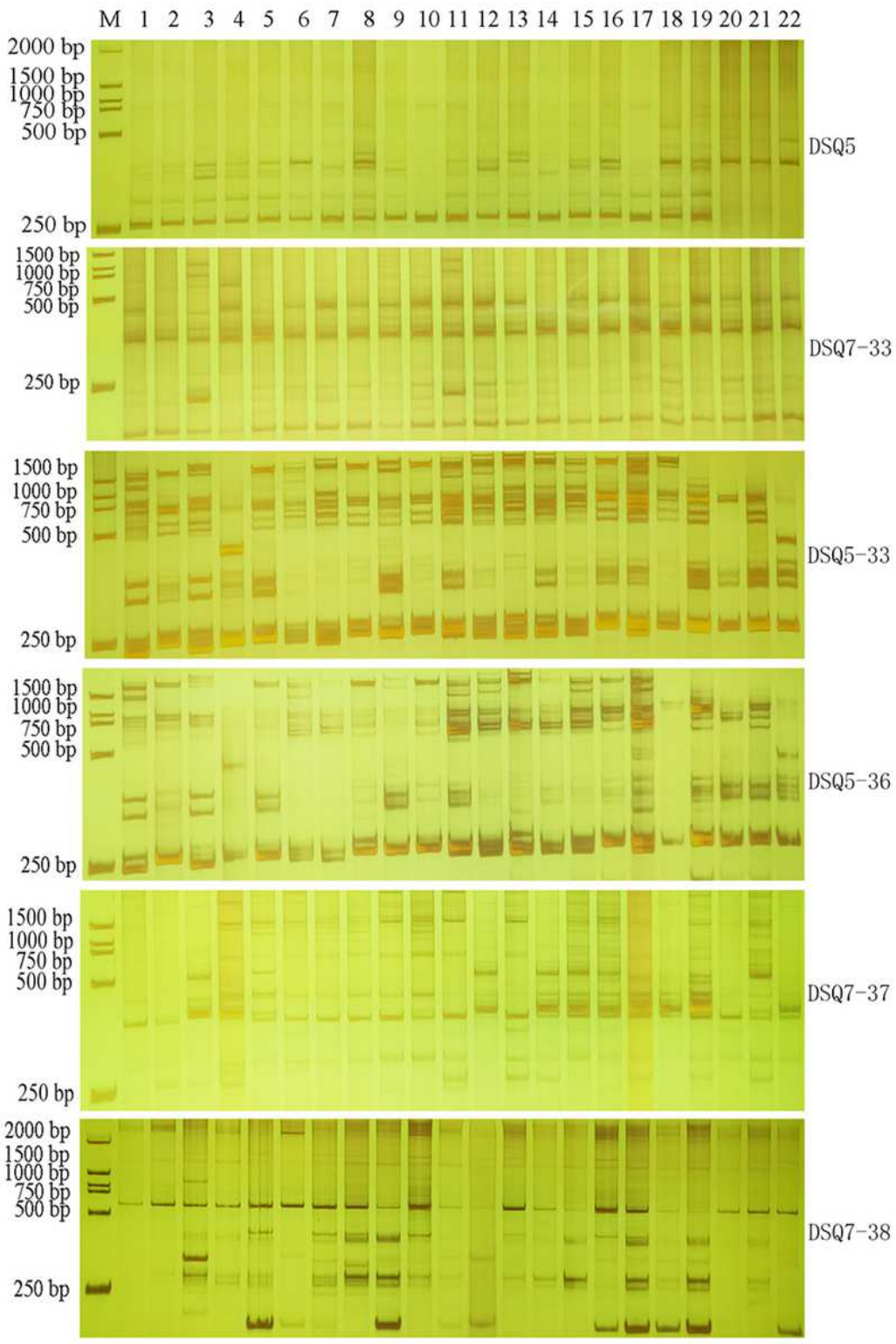

Figure 1

Amplification results of 6 polymorphism primers

Figure 2

Circular clustering tree. A 47 mixed materials. B 32 cultivated materials.C 15 wild materials

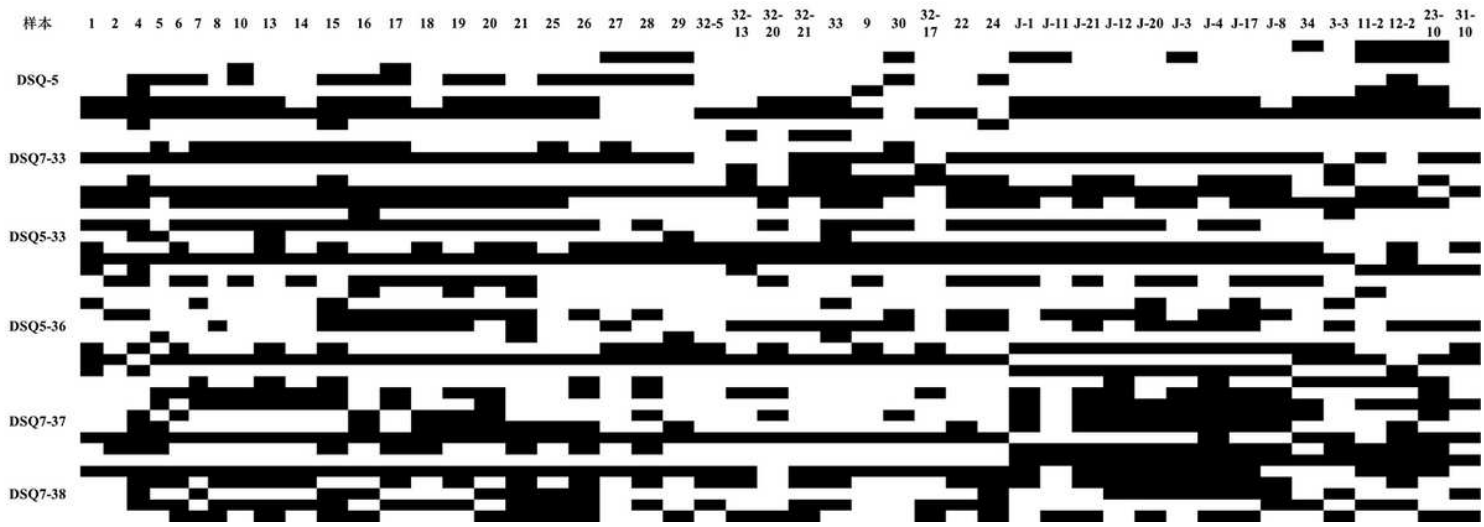

Figure 3

DNA fingerprints of 47 mixed materials

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation2Partialprimerinformation.docx