Genetic Diversity and Structure of Core Collection of Huangqi (Astragalus) Developed by Genomic Simple Sequence Repeat Markers

Fanshu Gong
Shanxi Agricultural University https://orcid.org/0000-0001-7556-7368
Yaping Geng
Shanxi Agricultural University
Pengfei Zhang
Shanxi Agricultural University
Feng Zhang
Shanxi Agricultural University
Xinfeng Fan
Shanxi Agricultural University
Yaling Liu (lylzpf@126.com)
Shanxi Agricultural University

Research Article

Keywords: Huangqi, Astragalus, SSR, core collection, molecular marker

Posted Date: January 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1218329/v1

Abstract

Huangqi (*Astragalus*) is a versatile herb that possesses several therapeutic effects against a variety of diseases, especially lung diseases. The aim of this study was to establish a core collection of *Astragalus* germplasm resources based on molecular 10 SSR markers. Based on 380 samples of *Astragalus* collected from different areas, five different methods were utilized to construct the core collection of *Astragalus*, including PowerCore-based M strategy, CoreFinder-based M strategy, Core Hunter-based stepwise sampling, PowerMarker-based simulated annealing algorithm based on allele maximization, and PowerMarker-based simulated annealing algorithm based on maximizing genetic diversity. Of the constructed *Astragalus* core collections, the CoreFinder-based M strategy was found to be the most suitable approach as it reserved all the alleles and most of the genetic diversity parameters were higher than those of the initial collection. Additional analyses demonstrated that the genetic diversity of the core collection matched the properties of the initial collection. Further, the phylogenetic trees indicated that the population structure of the core collection was similar to that of the initial collection. In addition, our results showed that the optimal grouping value of *K* was 2. The construction of a core collection is beneficial for the understanding, management, and utilization of *Astragalus*. Moreover, this study will act as a valuable reference for constructing core collections for other plants or fungi.

Introduction

The increase in plant genetic resources has significantly facilitated the development of agriculture, especially crop breeding (Escribano et al., 2008). In particular, a series of germplasm collections have provided genetic basis for crop improvement breeding. However, the long-term conservation of these collections is difficult (Escribano et al., 2008). The concept of a core collection was first proposed in 1984 (Frankel and Brown, 1984). A core collection is a minimum subset of initial collections that contains almost every possible genetic diversity (Frankel and Brown, 1984; Brown, 1989) The construction of a core collection is vital for the long-term conservation of genetic information, efficient management of initial collections, and the frequent utilization of germplasm (Liu et al., 2018). To date, core collections of important economic and staple crops have been constructed, including rice (*Oryza sativa*) (Zhao et al., 2010), maize (*Zea mays*) (Li et al., 2005), cotton (*Gossypium* spp.) (Ma et al., 2018; Hinze et al., 2015), wheat (*Triticum asetivum*) (Hao et al., 2006; Balfourier et al., 2007), soybean (*Glycine max*) (Oliveira et al., 2010), and winter mushroom (*Flammulina velutipes*) (Liu et al., 2018).

A core collection is conventionally established on the basis of agronomic and morphological characteristics via a variety of strategies, such as the random sampling strategy, constant allocation strategy, proportional allocation strategy, and the logarithm strategy (Brown, 1989; Ortiz et al., 1998; Holbrook et al., 1993). As most agronomic and morphological characteristics are easily impacted by environmental aspects, the genetic diversity of germplasm collections could not be directly represented using phenotype data (Hu et al., 2000; Xiao et al., 2010; Zhang et al., 2012). Thus, molecular markers have been considered to represent the genetic diversity of germplasms at the DNA sequence level (Zane et al., 2002). Among these markers, simple sequence repeats (SSRs; synonymy microsatellites) are tandemly arranged repeats of 1 – 6 base pair DNA sequences (Cheng et al., 2016). SSRs display a number of suitable characteristics for markers, such as co-dominant transmission, high polymorphism, and ease of detection (Powell et al., 1996). In addition, it has been reported that SSRs are abundant and ubiquitous in prokaryotic and eukaryotic genomes (Mrázek et al., 2007; Silver and Lee, 1992). Consequently, SSRs have become the most popular marker for plant geneticists and breeders.

Huangqi (*Astragalus*) is a versatile herb used in traditional Chinese medicines, and has been utilized for over 2000 years (Chu et al., 2010). Clinical applications have verified that *Astragalus* has therapeutic effects against heart disease, fatigue, hepatitis, and especially spleen and lung diseases (Gong et al., 2018; Chen et al., 2020). *Astragalus* is predominantly grown in China, and contains at least 278 species, 2 subspecies, 35 varieties, and 2 forma (Gong et al., 2018; Chen et al., 2020; Ma et al., 2002). Among these, *Astragalus membranaceus* (Fisch.) and *A. membranaceus* (Fisch.) var. *mongholicus* (Bunge) have been documented in the Chinese Pharmacopoeia and are mainly cultured in the north and northeast parts of mainland China (Gong et al., 2018; Chen et al., 2020; Ma et al., 2002). With the discovery of medical value, an increasing number of *Astragalus* germplasm resources are being uncovered. Thus, it is necessary to explore the potential functions and therapeutic effects, as well as the adverse efficient management and utilization impacts of *Astragalus*. Additionally, the construction of a core *Astragalus* collection could not only improve its management and utilization efficiency but also preserve the major genetic information of an entire germplasm collection. Hence, the establishment of a core collection of *Astragalus* is necessary.

Therefore, in this study, we constructed a core collection of 380 *Astragalus* varieties using five different methods. Among these methods, the CoreFinder-based M strategy was found to be the most suitable. Our results demonstrate that the core collection contained all the alleles and that most of the genetic diversity parameters were higher than the corresponding parameters in the initial collection. Moreover, our data indicated that the genetic diversity and structure of a core collection could represent the corresponding properties of an initial collection.

Materials And Methods

Plant materials

A total of 380 samples of *Astragalus* were collected from different areas, including 285 samples of *A. membranaceus* var. *mongholicus* and 95 samples of *A. membranaceus*. The single plant GPS system was utilized to determine the location of each individual plant (Supplemental Table 1). The samples in our study were obtained from different areas and no permission was necessary. All samples were identified by Professor Yaling Liu. The seeds of plants from each population were preserved in the herbarium of the College of Life Sciences at Shanxi Agricultural University. The leaves of each sample were harvested randomly and the distance between different samples was more than 10 m. The collected leaves were packaged in self-sealing bags filled with silica gel for DNA extraction. Experimental research on plants, including collection of plant material, was in accordance with guidelines of Shanxi Agricultural University.

DNA isolation and SSR genotyping

The global DNA genomes of all the 380 plant samples were isolated using the cetyltrimethylammonium bromide (CTAB) method (Springer and Nathan, 2010). Briefly, the samples were ground into fine powder using a mortar with liquid nitrogen. Then, the powder was dissolved with a CTAB buffer and incubated at 65 °C for 45 min. Next, phenol chloroform isoamyl alcohol (V/V/V=25:24:1) was used to exhaust the protein impurity. Finally, the DNA genome was dissolved with water or Tris-EDTA buffer after washing with 70% ethanol twice. After isolation, a nucleic acid protein detector and agarose gel electrophoresis were used to determine the concentration and purity of the DNA genome. Then, the concentration of each DNA genome was adjusted to 50 ng·μL^{-1} and stored at -20 °C.

A total of 10 pairs of SSR primers with high polymorphism and good reproducibility were selected to amplify the DNA genomes of the 380 *Astragalus* samples (Liu et al., 2019; Liu et al., 2014). The PCR products were detected using Fragment Analyzer^{TM} automatic capillary electrophoresis, which were then visualized with PROSize software.

Construction of core collections

Core collections were constructed using the following methods: (1) PowerCore V1.0 software operation construction based on the maximization strategy (M strategy, allele maximization) and heuristic algorithm (Kim et al., 2007). For sample selection, PowerCore V1.0 uses an advanced M strategy and heuristic algorithm to identify the optimal path from the initial stage to the final stage (Kim et al., 2007). The specific process is clicking the "first step" button to obtain the frequency distribution of each marker allele, then clicking the "second step" to enter the core collection interface, and finally clicking "run" to obtain a core set; (2) CoreFinder V. 1.1 software operation construction based on M strategy and the use of Las Vegas-style random algorithm (Cipriani et al., 2010). These two methods (PowerCore V1.0 and CoreFinder V. 1.1 software) did not need to set a sampling ratio during the core collection construction process, as a reasonable sampling ratio was generated automatically; (3) Stepwise extraction method based on genetic distance (Core Hunter software). Core Hunter v. 2.0 is a fast core collection selection program based on multiple genetic diversity measures and a hybrid copy search algorithm (Beukelaer et al., 2012). This software allows the selection of a sampling ratio and certain genetic metrics as selection criteria; (4) Simulated annealing algorithm based on allele maximization (SANA); (5) Simulated annealing algorithm based on maximizing genetic diversity (SAGD). These two methods were conducted using the core set section of the PowerMarker software (Liu et al., 2005). By setting the sampling ratio the same as PowerCore and PowerMarker, the gene frequencies of all the germplasms were detected using the corresponding methods (SANA and SAGD, respectively).

Data analysis

GenAlEx 6.5 software was used for the population genetic analyses, which includes frequency- and distance-based analyses (Peakall and Smouse, 2012). The genetic diversity index was analyzed using GenAlEx 6.503 software, including the average number of alleles, average effective number of alleles (*Ne*), average Shannon information index (*I*), average Nei's genetic diversity index (*H*), average observed heterozygosity (*Ho*), and the average expected heterozygosity (*He*). To verify the effectiveness of the core collection, a *t* test was performed on the genetic diversity parameters of the initial collection and each core collection using SPSS 23.0 software. Further, a principal coordinate analysis (PCoA) was conducted to evaluate the core collection using GenAlEx 6.503 software.

After the construction of the core collection, the Nei's genetic distance of different medicinal *Astragalus* was calculated using PowerMarker software (Liu et al., 2005). Moreover, the phylogenetic tree based on the NJ adjacency clustering was constructed using MEGA 7.0 software (Kumar et al., 2016; Saitou and Nei, 1987). To explore the genetic structures of the initial and core collections of *Astragalus*, a Bayesian cluster analysis was performed with STRUCTURE software (Pritchard et al., 2000). Briefly, the number of groups (K) was set to 1 – 10 and each K value was simulated 10 times. The Markov chain Monte Carlo (MCMC) at the beginning of the burn-in-period was set to 10,000 and the MCMC after the burn-in-period was 100,000 times. Finally, the STRUCTURE data was subjected to the Structure Harvester website to determine the optimal K value.

Results

Core collection of Astragalus

In this study, a core collection of 380 *Astragalus* germplasm resources (consisting of 95 *A. membranaceus* and 285 *A. mongholicus*) was constructed using 5 different methods based on SSR molecular markers. Based on the M strategy, the results of the PowerCore and CoreFinder analyses showed that the core collection retained 36% and 45% of the original germplasm resources, respectively. Briefly, 100 *A. mongolicus* and 45 *A. membranaceus* germplasms were selected by the PowerCore software, whereas 93 *mongolicus* and 42 *A. membranaceus* germplasms were retained by the CoreFinder software (Table 1).Regarding the core collection established by stepwise sampling, 100 *A. mongolicus* and 45 *A. membranaceus* germplasms were selected (Table 1). Based on the PowerMarker software, 116 *mongolicus*, and 29 *A. membranaceus* germplasm were selected by the SANA method, and 110 *mongolicus* and 35 *A. membranaceus* germplasm were selected by the simulated annealing algorithm based on maximizing genetic diversity (SAGD) method (Table 1).

Table 1
The construction of core collection by 5 different methods.

Variations	Locations	Serial number	Count of core collections (Number)			
			M strategy[1] (PowerCore)	M strategy (CoreFinder)	Stepwise sampling (Core Hunter)	SANA (Pow
A. mongholicus	Hongshaba Village, Xiashihao Township, Guyang County, Baotou City	GHSB	5 (2,4,7,8,14)	5 (2,4,7,8,14)	5 (3,7,12,13,20)	9 (4,5,7
	Hadamen Forest Park, Daqingshan Township, Wuchuan County, Hohhot City	GSLY	8(3,7,12,14,15,16,18,20)	9 (3,7,10,12,14,15,16,18,20)	6 (1,2,3,6,10,19)	6 (4,5,7
	Huanggangliang Forest Farm, Xinjing Township, Keshiketeng Banner, Chifeng City	GHGL	10 (2,5,8,12,13,15,16,17,18,19)	7 (2,5,12,13,16,17,18,19)	7 (11,12,13,14,15,17,19)	5 (5,10
	Huangtuchang Village, Xiashihao Township, Guyang County, Baotou City	GHTC	4(7,8,10,19)	4 (7,8,10,19)	6 (4,5,8,14,16,19)	10 (1,5,7
	Wanlongquan Village, Qiaotou Town, Wengniute Banner, Chifeng City	GWLQ	6(2,3,8,12,13,20)	6(2,3,8,12,13,20)	6 (2,3,7,9,11,18	11 (1,4,5
	Heichengzi, Zhenglan Banner, Xilin Gol City	GHCZ	3 (3,5,8)	3 (3,5,8)	3 (5,14,15	5 (1,10
	Wuyi Village, Dakulian Township, Xinghe County, Ulanqab City	GWYC	10 (2,3,4,6,8,11,12,16,1719)	10 (2,3,4,6,8,11,12,16,1719)	4 (1,3,7,10)	3 (1,9,1
	Manzhouli, Inner Mongolia	GMZL	3 (3,5,8)	3 (3,5,8)	2 (4,6)	5 (1,3,5
	Pangquangou, Shanxi Province	GPQG	1 (2)	3 (1,2,3)	2 (1,4)	3 (1,2,4
	Pangjiayao Village, Yanggao County, Shanxi Province	GPJY	8 (1, 3,5,7,11,13,16,17)	7 (1,3,5,7,11,13,17)	7 (2,5,9,10,12,13,15)	8 (2,4,9

Note: [1] M strategy, Maximization strategy.

[2] SANA, Simulated annealing algorithm maximizing the number of alleles.

[3] SAGD, Simulated annealing algorithm maximizing the genetic diversity.

Variations	Locations	Serial number	Count of core collections (Number)			
			M strategy[1] (PowerCore)	M strategy (CoreFinder)	Stepwise sampling (Core Hunter)	SANA (Pow
	Shijiaping Village, Guangling County, Shanxi Province	GSJP	5 (2,4,5,7,13)	3 (2,4,5)	11 (3,4,7,8,9,10,12,13,14,17,19	6 (4,5,6
	Lucaowan Village, Tianzhen County, Shanxi Province	GLCW	5 (1,2,3,4,16)	5 (1,2,3,4,16)	4 (1,2,6,9)	8 (2,3,5
	Pangjiatao Village, Baimashi Township, Ying County, Shanxi Province	GPJT	7 (4,5,7,9,14,16,18)	6 (4,7,9,14,16,18)	8 (1,3,4,8,11,13,15,17)	7 (1,4,5
	Zeqingling Village, Hunyuan County, Shanxi Province	GZQL	8 (2,6,7,12,16,17,19,20)	7 (2,6,7,12,16,17,20)	6 (6,8,13,15,16,19)	10 (3,5,6
	Yangci Village, Dai County, Shanxi Province	GDXY	9 (1,2,4,5,10,11,13,17,19)	7 (1,4,5,10,11,13,19)	9 (1,2,3,5,9,11,14,18,19)	5 (1,5,1
	Xinxingtun, Jilin City	GXXT	5 (7,8,9,12,15)	5 (7,8,9,12,15)	10 (1,2,5,7,8,11,14,15,18,19	7 (3,4,6
	Hadamen Township, Jilin City	GHDM	3 (1,3,6)	3 (1,3,6)	1 (5)	2 (5,6)
Total			**100**	**93**	**100**	**116**
A. membranaceus	Xinxing Township Forest Farm, Jilin City	MJLC	9 (1,2,6,8,9,11,12,15,17)	7 (1,2,6,8,9,11,12)	8 (7,8,10,11,14,15,16,18)	3 (14,1
	Zeqingling Village, Hunyuan County, Shanxi Province	MSHY	11 (1,2,3,7,8,9,10,11,13,16, 18)	10 (1,2,3,8,9,10,11,13,16, 18)	12 (1,2,3,4,6,8,9,10,12,17,18,19)	10 (1,2,3
	Hailin City, Heilongjiang Province	MHLJ	13 (1,2,4,5,7,9,10,11,12,13,14,18,19)	12 (1,4,6,7,9,10,11,12,13,14,18,19)	8 (2,4,6,7,10,12,15,19)	8 (3,8,1
	Xinbin County, Liaoning Province	MLXB	5 (4,8,9,11, 17)	6 (2,4,8,9,11,17)	6 (3,7,8,10,15,18)	7 (6,11
	Qingyuan County, Liaoning Province	MLQY	7 (1,4,6,8,12,16,18	7 (1,4,6,8,12,13,18	9 (1,2,3,5,6,8,11,14,15	7 (1,7,8
Total			**45**	**42**	**45**	**29**

Note: [1] M strategy, Maximization strategy.

[2] SANA, Simulated annealing algorithm maximizing the number of alleles.

[3] SAGD, Simulated annealing algorithm maximizing the genetic diversity.

Genetic diversity of the Astragalus **core collection**

The genetic diversity parameters of the *Astragalus* samples are summarized in Table 2. The results show that 100% of the alleles were selected by the PowerCore and CoreFinder software. Moreover, the *N*, *H*, *I*, and *He* values of the core collections obtained using PowerCore and CoreFinder software were higher than those of the initial collection, with the exception of *Ho*. In addition, the retention ratios of *Ne*, *H*, *I*, *He* were more than 100% and those of *Ho* were 96.137% and 97.425%, respectively, for the PowerCore and CoreFinder software. The *t*-test data revealed that all the genetic diversity parameters were not significantly different.

Table 2
Analysis of the *Astragalus* core collection data.

	Sample number	Average number of alleles ± SD (*Na*)	Average number of effective alleles±SD (*Ne*)	Average Nei's genetic diversity index±SD (*H*)	Average Shannon Information Index±SD (*I*)	Average observed heterozygosity±SD (*Ho*)	Average expected heterozygosity±SD (*He*)
Initial collection	380	33.700±1.770	6.843±1.012	0.831±0.019	2.365±0.115	0.466±0.100	0.832±0.019
Core collection (PowerCore)	145	33.700±1.770	9.162±1.284	0.875±0.014	2.679±0.102	0.448±0.093	0.878±0.014
Sampling ratio (%)	38%	100	133.889	105.295	113.277	96.137	105.529
t-value[1]		0.000	-1.419	-1.859	-2.038	0.132	-0.238
p-value		1.000	0.173	0.079	0.056	0.897	0.814
Core Collection (CoreFinder)	136	33.700±1.770	8.845±1.226	0.870±0.015	2.656±0.105	0.454±0.094	0.874±0.015
Sampling ratio (%)	36%	100	129.256	104.693	112.304	97.425	105.048
t-value		0.000	-1.2593	-1.6336	-1.8748	0.0832	-1.7214
p-value		1.000	0.2240	0.1197	0.0771	0.9346	0.1023
Core Collection (Core Hunter)	145	24.800±1.781	6.725±0.930	0.830±0.018	2.306±0.110	0.469±0.100	0.833±0.018
Sampling ratio (%)	38%	73.591	98.276	99.880	97.505	100.644	100.120
t-value		3.5437	0.0857	0.0150	0.3714	-0.0261	-0.0562
p-value		0.0023	0.9326	0.9882	0.7147	0.9795	0.9558
Core Collection (SANA[2])	145	24.700±1.391	6.620±0.779	0.833±0.016	2.304±0.095	0.454±0.103	0.836±0.016
Sampling ratio (%)	38%	73.294	96.741	100.241	97.421	97.425	100.481
t-value		3.9975	0.1747	-0.1007	0.4056	0.0802	-0.1730
p-value		0.0008	0.8633	0.9209	0.6898	0.9370	0.8646
Core Collection (SAGD[3])	145	24.500±1.939	6.762±1.018	0.829±0.019	2.303±0.116	0.464±0.104	0.832±0.019
Sampling ratio (%)	38%	72.700	98.816	99.760	97.378	99.571	100
t-value		3.5035	0.0563	0.0523	0.3787	0.0118	-0.0149
p-value		0.0025	0.9557	0.9589	0.7094	0.9907	0.9883

Note: [1] $t_{0.05}$=1.960, $t_{0.01}$=2.576.

[2] SANA, Simulated annealing algorithm maximizing the number of alleles.

[3] SAGD, Simulated annealing algorithm maximizing the genetic diversity.

Similarly, the core collection established using Core Hunter based on the stepwise clustering method retained approximately 74% of the alleles. Further, the allele retention ratios of the SANA and SAGD methods using PowerMarker were 73.294% and 72.700%, respectively. In addition, the *t*-test data showed that the difference between the core collections selected by these three methods and the initial collection were extremely significant ($p < 0.01$). Thus, the genetic diversity parameters (*Ne*, *H*, *I*, *Ho*, *He*) of the obtained and initial collections were further explored. The analysis results showed that the initial collection possessed a slightly greater genetic diversity than most of the obtained core collections. In addition, the statistical data showed that they were not significantly different.

PCoA of Astragalus core collection

To identify whether the genetic diversity of the core collection constructed by these 5 different methods was consistent with the genetic diversity of the initial collection, a PCoA was performed using the GeneAlEx software. The results showed that the two-dimensional coordinate distributions of the core and initial collections were similar (Fig. 1). Moreover, the distributions of *A. membranaceus* and *A. mongholicus* were distinct (Fig. 1). On this basis, the genetic distribution of the core collection could represent the genetic location of the initial collection.

Comparison of the Astragalus core collections established using different methods

Based on the detailed analysis of the core collections with different construction methods (Table 2), the core collections based on the M strategy, which were conducted using PowerCore and CoreFinder software, reserved the most alleles (100% retention). This indicates that the core collection retained all the alleles that were detected at the initial collection. However, approximately 27% of alleles were missed in the selection processes of the other three methods. Moreover, the core collection constructed by CoreFinder only contained 36% (135/380) of the original germplasms, which was lower than the proportion of the core collection selected by the other methods. As listed in Table 2, 33% (93/285) and 44% (42/95) of the *A. mongolicus* and *A. membranaceus* germplasms, respectively, were reserved by this software. Not including *Ho*, the genetic diversity parameters (*Ne*, *H*, *I*, *Ho*, and *He*) of the M strategy-constructed core collections were higher than those constructed by the other three methods. Specifically, the *Ne*, *H*, *I*, *Ho*, and *He* values of the CoreFinder-established core set were 0.870 ± 0.015, 2.656 ± 0.105, 0.454 ± 0.094, and 0.874 ± 0.015, respectively. The retention proportions, as compared with the initial collection, were 104.693%, 112.304%, 97.425%, and 105.048%, respectively. In addition, the results of the *t*-test verified that the difference between the core and initial collections was not significant ($p > 0.05$). Thus, the M strategy-based CoreFinder method is the most suitable for establishing a core collection of *Astragalus* germplasms.

Genetic structure of the Astragalus core collection

To explore the phylogenetic relationships of the initial and core collections, NJ clustering were performed using MEGA7.0 software. According to the alignment results, the phylogenetic tree discovered that the initial germplasms could be divided into three groups. As shown in Figs. 2 and 3, the distributions of the three groups were similar, suggesting that the population structures of the initial germplasm resources and the core collection were similar. These results further support that the core collection constructed by CoreFinder could represent the initial collection.

To investigate the population structure of the core collection and their optimal *K* value, the data was subjected to the STRUCTURE software (Fig. 4). The results show that the optimal grouping number of *K* was 2 when the Δ*K* value reached the maximum value (ΔK = 1027.95) (Figs. 4D and 4E). Fig. 4E also shows that the core collection was divided into two groups (shaded in red and green), wherein the red and green groups consisted of the *A. mongolicus* and *A. membranaceus* populations, respectively, which was in consistent with the phylogenetic systematics. Thus, the population structure of the core collection conformed to the initial population structure.

Molecular variance analysis of core collection

To investigate the genetic diversity of the germplasm collections, molecular variance analyses (AMOVAs) of the initial and core collections were performed. The results showed that genetic variation mainly developed within the collections rather than among the collections (Table 3). In addition, the genetic variation in the initial and core collections mainly occurred with the population, occupying 85% and 89% of the genetic variance, respectively (Table 3, Fig. 5). The genetic variation between the two *Astragalus* variations occupied 10% of the variance of the initial collection and 7% of the variance of the core collection (Table 3, Fig. 5). Moreover, the genetic variation between the groups accounted for 6% and 4% of the total variation in the initial and core collections (Table 3, Fig. 5), respectively. In addition, the distributions of the genetic variation in the initial and core collections were similar. This may further verify that the genetic diversity between the initial and core collections were similar.

Table 3
Molecular variance analysis between the initial collection and the core collection.

	df[1]	SS[2]	MS[3]	Est. Var.[4]	Percentage of total variation
Among Collections	1	6.657	6.657	0.006	0%
Within Collections	1030	4388.961	4.261	4.261	100%
Total	1031	4395.618		4.267	100%
Core Collection					
Among Species	1	45.057	45.057	0.312	7%
Among Pops	20	162.601	8.130	0.168	4%
Among Indiv	114	695.250	6.099	1.976	43%
Within Indiv	136	292.000	2.147	2.147	46%
Total	271	1194.908		4.603	100%
Initial Collection					
Among Species	1	137.386	137.386	0.427	10%
Among Pops	20	297.441	14.872	0.278	6%
Among Indiv	358	1914.225	5.347	1.562	35%
Within Indiv	380	845.000	2.224	2.224	50%
Total	759	3194.053		4.490	100%
Note: [1] df, degrees of freedom.					
[2] SS, sum of squares.					
[3] MS, mean squares.					
[4] Est. Var., Estimated variance.					

Discussion

With the rapid development of biotechnology, numerous germplasm resources have been identified by plant scientists and breeders, making the efficient management and utilization of these collections necessary. The construction of a core collection based on molecular markers, especially SSR markers, was found to be the optimal approach (Brown, 1989). Huangqi (*Astragalus*) is a versatile herb used for medical applications, including heart disease, fatigue, hepatitis, and especially spleen and lung diseases (Gong et al., 2018; Chen et al., 2020). Although the core collections of many plants and edible mushrooms have been established, such as rice, wheat, cotton, winter mushroom, and shiitake (*Lentinula edodes*) (Liu et al., 2018; Ma et al., 2018; Hinze et al., 2015; Hao et al., 2006; Xiao et al., 2010; Tiwari et al., 2015), the construction of a core collection of *Astragalus* has not been conducted. Therefore, in this study, we successfully established a core collection of *Astragalus* with 100% allelic representation using 36% of the germplasm proportion (33% for *A. mongolicus* and 44% for *A. membranaceus*) based on 10 SSR markers. In addition, the genetic diversity of the core collection was found to be suitable for representing the entire collection. Moreover, the selected germplasms in the core collection represented two cultivated variations, including *A. mongolicus* and *A. membranaceus*. It was further proven that the population structures of the initial and core collections were similar.

In recent decades, various molecular markers have been developed, but SSR markers have become dominant because of their practical characteristics (Zhang et al., 2012). In this work, we further verified that SSR markers are suitable by constructing an *Astragalus* core collection. Using CoreFinder software, we obtained a suitable core collection based on SSR markers. In this study, we compared five different methods to construct a core collection of 380 *Astragalus* samples. In addition, four different algorithms (M strategy, stepwise sampling, SANA, and SAGD) were utilized with four different software programs for construction (PowerCore, CoreFinder, Core Hunter, and PowerMarker). Our results showed that CoreFinder-based M strategy was the most suitable for constructing the *Astragalus* core collection. Although this traditional Chinese herb possesses various therapeutic and health-care functions, the germplasm resources are numerous and intricate. Therefore, this study suggests that we could use SSR markers with the help of several analyzing software programs to manage and utilize the complicated herbal germplasm.

However, the genetic diversity of the germplasm resources could not be directly represented by the phenotypic data, but molecular markers are considered to be a manifestation of the genetic diversity of the germplasm collections. Additionally, the agronomic and morphological properties of crops are highly vulnerable to the environmental variations (Hu et al., 2000; Xiao et al., 2010; Zhang et al., 2012; Zane et al., 2002). Thus, the core collection based on 380 *Astragalus* germplasm samples efficiently preserved numerous agronomic and morphological properties of *Astragalus*, which will promote the efficient management and fast utilization of the *Astragalus* germplasm, consequently facilitating its breeding development. Overall, this study not only successfully established a core collection of *Astragalus* germplasm but also furthered *Astragalus* research, as well as Chinese herb research.

In this study, a core collection of 380 *Astragalus* samples was successfully constructed. As compared with other methods, CoreFinder software, in combination with the M strategy, was found to be the most suitable approach for constructing the *Astragalus* core collection. Further, we verified that the genetic diversity and structure of this core collection represented that of the initial germplasm resources. The constructed practical core collection of *Astragalus* will act as a useful reference for the characterization, management, and utilization of *Astragalus*, and provide novel insights for other herbs or crops.

Declarations

Acknowledgements The authors are thankful to each of the local staff and planting bases that assisted the research team in taking samples, as well as the good scientific research environment provided by the school.

Authors' contributions PZ and YL designed the experiment, analyzed the data and completed the manuscript. YG and FG participated in writing, and discussed the results. YG, FG, FX and ZF carried the experiments. All authors have read and approved the manuscript. This work was supported by the National Natural Science Foundation of China (32070378), the Natural Science Foundation of Shanxi Province of China (201901D111217), Biological breeding project of Shanxi Agricultural University (YZGC136). The funder provided the research fund and no role in the experimental design, data analysis, decision to publish, or preparation of the manuscript.

Funding This work was supported by the National Natural Science Foundation of China (32070378), the Natural Science Foundation of Shanxi Province of China (201901D111217), Biological breeding project of Shanxi Agricultural University (YZGC136). The funder provided the research fund and no role in the experimental design, data analysis, decision to publish, or preparation of the manuscript.

Availability of data and material All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflict of interest The authors declare no conflict of interest.

References

1. Balfourier F, Roussel V, Strelchenko P, Exbrayat-Vinson F, Sourdille P, Boutet G, Koenig J, Ravel C, Mitrofanova O, Beckert M (*2007*) *A worldwide bread wheat core collection arrayed in a 384-well plate.Theoretical & Applied Genetics114*:*1265-1275*
2. Beukelae HD, Smykal P, Davenport GF, Fack V (*2012*). *Core hunter II: fast core subset selection based on multiple genetic diversity measures using mixed replica search.BMC Bioinformatics13*:*312-312*
3. Brown HdA (*1989*) *Core collections: a practical approach to genetic resources management.Genome31*:*818-824*
4. Chen Z, Liu L, Gao C, Chen W, Vong CT, Yao P, Yang Y, Li X, Tang X, Wang S, Wang Y (*2020*) *Astragali Radix (Huangqi): A promising edible immunomodulatory bal medicine.Journal of Ethnopharmacology258*:*112895*
5. Chu C, Lian WQ, Liu EH, Bin L, Wen G, Ping L (*2010*) *Radix Astragali (Astragalus): Latest Advancements and Trends in Chemistry, Analysis, Pharmacology and Pharmacokinetics.Current Organic Chemistry14*:*1792-1807*
6. Cipriani G, Spadotto A, Jurman I, Gaspero DG, Crespan M, Meneghetti S, Frare E, Vignani R, Cresti M, Morgante M (*2010*). *The SSR-based molecular profile of 1005 grapevine (Vitis vinifera L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin.Theoretical and Applied Genetics121*:*1569-1585*
7. Escribano P, Viruel MA, Hormaza JI (*2008*) *Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya (Annona cherimola, Annonaceae), an underutilised subtropical fruit tree species.Annals of Applied Biology153*:*25-32*
8. Frankel O, Brown A (*1984) Current plant genetic resources–a critical appraisal. iv.oxford & ibh publ.co. DOI*: http://dx.doi.org/
9. Gong AGW, Duan R, Wang HY, Kong XP, Dong TTX, Tsim KWK, Chan K (*2018*) *Evaluation of the Pharmaceutical Properties and Value of Astragali Radix.Medicines (Basel)5*:*46*
10. Hao CY, Zhang XY, Wan LF, Dong YS, Shang XW, Jia JZ (*2006*) *Genetic diversity and core collection evaluations in common wheat germplasm from the Northwestern Spring Wheat Region in China.Molecular Breeding17*:*69-77*
11. Hintum TJL, Bothmer R, Visser DL (*2010*) *Sampling strategies for composing a core collection of cultivated barley (hordeum vulgare s. iat.) collected in china.Hereditas122*:*7-17*
12. Hinze LL, Fang DD, Gore MA, Scheffler BE, Yu JZ, Frelichowski J, Percy RG (*2015*). *Molecular characterization of the Gossypium Diversity Reference Set of the US National Cotton Germplasm Collection. TAG Theoretical and applied geneticsTheoretische und angewandte Genetik128*:*313-327*
13. Holbrook CC, Anderson WF, Pittman RN (*1993*). *Selection of a core collection from the u.s. germplasm collection of peanut.Crop Science68*:*6392-9*
14. Hu J, Zhu J, Xu HM (*2000*) *Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops.Theoretical & Applied Genetics101*:*264-268*
15. Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG, Park YJ (*2007*) *PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets.Bioinformatics (Oxford, England)23*:*2155-2162*
16. Kumar S, Stecher G, Tamura K (*2016*) *MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.Molecular biology and evolution33*:*1870-1874*
17. Liu K, Muse SV (*2005*) *PowerMarker: an integrated analysis environment for genetic marker analysis.Bioinformatics (Oxford, England)21*:*2128-2129*

18. Liu XB, Li J, Yang ZL (*2018*) *Genetic diversity and structure of core collection of winter mushroom (Flammulina velutipes) developed by genomic SSR markers.Hereditas155*:*3*
19. Liu YL, Geng YP, Xie XD, Wang F, Zhang PF (*2019*) *Genetic Diversity and Genetic Structure Analysis of Astragalus Based on SSR Molecular Marker.Acta Agrestia Sinica27*:*1153-1162*
20. Liu YL, Wang WQ, Hou JL, Zhang R, Yang WX, Liu FB, Wei SL (*2014*) *The optimization and primary application for SSR-PCR Reaction System of the Astragalus.Lishizhen Medicine and Materia Medica Research25*:*2227-2229*
21. Ma XQ, Shi Q, Duan JA, Dong TT, Tsim KW (*2002*) *Chemical analysis of Radix Astragali (Huangqi) in China: a comparison with its adulterants and seasonal variations.Journal of agricultural and food chemistry50*:*4861-4866*
22. Ma Z, He S, Wang X, Sun J, Zhang Y, Zhang G, Wu L, Li Z, Liu Z, Sun G (*2018*) *Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield.Nature genetics50*:*803-813*
23. Mrázek J, Guo X, Shah A (*2007) Simple sequence repeats in prokaryotic genomes. Proceedings of the National Academy of Sciences 104: 8472-8477.*
24. Oliveira MF, Nelson RL, Geraldi IO, Cruz CD, Toledo J (*2010*) *Establishing a soybean germplasm core collection.Field Crops Research119*:*277-289*
25. Ortiz R, Ruiz-Tapia EN, Mujica-Sanchez A (*1998*) *Sampling strategy for a core collection of peruvian quinoa germplasm.Theoretical and Applied Genetics96*:*475-483*
26. Peakall R, Smouse PE (*2012*) *GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update.Bioinformatics (Oxford, England)28*:*2537-2539*
27. Powell W, Machray GC, Provan J (*1996*) *Polymorphism revealed by simple sequence repeats.Trends in Plant Science1*:*215-222*
28. Pritchard JK, Stephens M, Donnelly P (*2000*) *Inference of population structure using multilocus genotype data.Genetics155*:*945-959*
29. Saitou N, Nei M (*1987*) *The neighbor-joining method: a new method for reconstructing phylogenetic trees.Molecular biology and evolution4*:*406-425*
30. Silver, Lee M (*1992*) *Bouncing off microsatellites.Nature Genetics2*:*8-9*
31. Springer NathanM (*2010*) *Isolation of plant dna for pcr and genotyping using organic extraction and ctab. Cold Spring Harbor Protocols 11: pdb.prot5515. DOI*: 10.1101/pdb.prot5515
32. Tiwari KK, Singh A, Pattnaik S, Sandhu M, Kaur S, Jain S, Tiwari S, Mehrotra S, Anumalla M, Samal R, Bhardwaj J, Dubey N, Sahu V, Kharshing GA, Zeliang PK, Sreenivasan K, Kumar P, Parida SK, Mithra SVA, Rai V, TyagiW, Agrawal PK, Rao AR, Pattanayak A, Girish (*2015*) *Identification of a diverse mini-core panel of I ndian rice germplasm based on genotyping using microsatellite markers.Plant Breeding134*:*164-171*
33. Xiao Y, Liu W, Dai Y, Fu C, Bian Y (*2010*) *Using SSR markers to evaluate the genetic diversity of Lentinula edodes' natural germplasm in China.World Journal of Microbiology & Biotechnology26*:*527-536*
34. Yu L, Shi Y, Cao Y, Wang T (*2005*) *Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data.Genetic Resources & Crop Evolution51*:*845-852*
35. Zane L, Bargelloni L, Patarnello T (*2010*) *Strategies for microsatellite isolation: a review.Molecular Ecology11*:*1-16*
36. Zhang RY, Hu DD, Gu JG, Hu QX, Zuo XM, Wang HX (*2012*) *Development of SSR markers for typing cultivars in the mushroom Auricularia auricula-judae.MYCOLOGICAL PROGRESS11*:*587-592*
37. Zhao W, Cho GT, Ma KH, Chung JW, Gwag JG, Park YJ (*2010*) *Development of an allele-mining set in rice using a heuristic algorithm and SSR genotype data with least redundancy for the post-genomic era.Molecular Breeding26*:*639-651*

Figures

Figure 1

Principal coordinate analysis (PCoA) of the *Astragalus* core collection. (A-E) PCoA data of the core collection constructed by Powercore, CoreFinder, Core Hunter, PowerMarker (SANA), and PowerMarker (SAGD), as compared with the initial collection, respectively. The solid symbols represent the selected core collection and the hollow symbols indicate the unselected core collection. The circular symbols represent the *A. mongolicus* core set and the quadrate symbols indicate the *A. membranaceus* core set.

Figure 2

Phylogenetic tree of the 380 *Astragalus* germplasm resources.

Figure 3

Phylogenetic tree of the core collection constructed by CoreFinder.

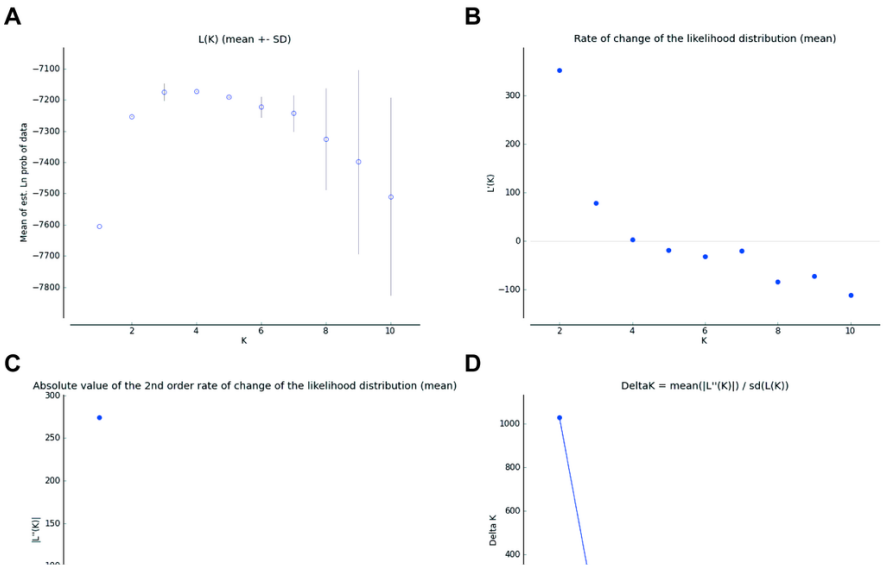

Figure 4

Gene structure of the core collection. (A) Estimated average likelihood (*K*) distribution (average + standard deviation) from 2 to 10 possible clusters. (B) Rate of change of likelihood distribution from 2 to 10 possible clusters. (C) Change rate of likelihood distribution from 2 to 10 possible clusters. (D) Δ*K* value distribution based on the change rate of L(*K*) between continuous *K* values. (E) Estimated population structure of the core collection for *K* = 2. The serial number is labeled beneath the Figure. Variations of *Astragalus* are labeled above the Figure. Red shading represents *A. mongolicus* and green shading indicates *A. membranaceus*.

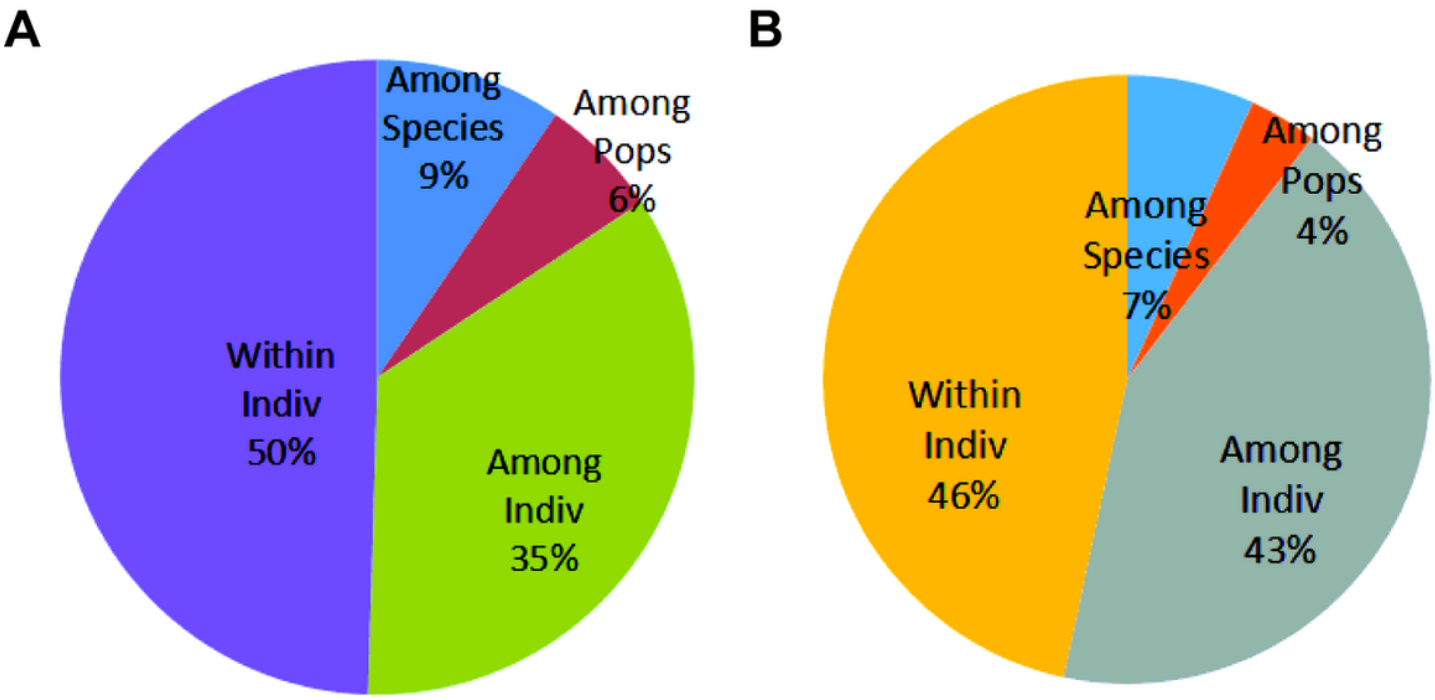

Figure 5

Molecular variance analysis of the initial and core collections. (A) Genetic variation of the (A) core collection and (B) initial collection.

Supplementary Files

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

- SupplementalTable1.docx