PLOS ONE

RESEARCH ARTICLE

Selection of a core collection of *Prunus sibirica* L. germplasm by a stepwise clustering method using simple sequence repeat markers

Yongqiang Sun, Shengjun Dong*, Quangang Liu, Jianhua Chen, Jingjing Pan, Jian Zhang

School of Forestry, Shenyang Agricultural University, Shenyang, China

* dsj928@163.com

OPEN ACCESS

Citation: Sun Y, Dong S, Liu Q, Chen J, Pan J, Zhang J (2021) Selection of a core collection of *Prunus sibirica* L. germplasm by a stepwise clustering method using simple sequence repeat markers. PLoS ONE 16(11): e0260097. https://doi.org/10.1371/journal.pone.0260097

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: May 6, 2021

Accepted: November 2, 2021

Published: November 19, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by National Key R&D Program Project of China-Germplasm Creation and Breeding of Almond-Apricot (Grand code:2019YFD1001203). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Prunus sibirica is an economically important tree species that occurs in arid and semi-arid regions of northern China. For this species, creation of a core collection is critical for future ecological and evolutionary studies, efficient economic utilization, and development and management of the broader collection of its germplasm resources. In this study, we sampled 158 accessions of *P. sibirica* from Russia and China using 30 pair of simple sequence repeat molecular markers and 30 different schemes to identify candidate core collections. The 30 schemes were based on combinations of two different sampling strategies, three genetic distances, and five different sample sizes of the complete germplasm resource. We determined the optimal core collection from among the 30 results based on maximization of genetic diversity among groups according to Number of observed alleles (N_a), Number of effective alleles (N_e), Shannon's information index (I), Polymorphic information content (PIC), Nei gene diversity (H) and compared to the initial collection of 158 accessions. We found that the optimal core collection resulted from preferred sampling at 25% with Nei & Li genetic distance these ratios of N_a, N_e, I, PIC and H to the complete 158 germplasm resources were 73.0%, 113%, 102%, 100% and 103%, respectively, indicating that the core collection comprised a robust representation of genetic diversity in *P. sibirica*. The proposed core collection will be valuable for future molecular breeding of this species and management of its germplasm resources.

Introduction

Siberian apricot (*Prunus sibirica*) is a shrub or tree species of the genus Rosaceae family. *P. sibirica* is widely distributed across a wide range of geographical areas in the northern regions of China and Russia [1]. Duo to its strong ecological adaptability, *P. sibirica* can be used as a pioneer tree species for improving the environment in semi-arid and arid areas [2]. Its kernels are of great economic value and have become an important industry for the people in the producing areas to get rid of poverty and become rich [3]. However, the cultivation of *P. sibirica* and its promotion within the horticultural and silvicultural industries still face challenges. In

Competing interests: The authors have declared that no competing interests exist.

particular, the cultivar that high-quality and high-yield remain scarce, but this can be overcome by plant breeding research, especially molecular breeding approaches, to expand the scale of cultivation of *P. sibirica*.

Plant breeding process needs to rely on large-scale germplasm resources. Thus, the collection, evaluation, management and utilization of germplasm resources are supported by many governments or organizations worldwide [4–7]. As of 2016, there are more than 1300 germplasm resources preservation banks around the world, and approximately 6.1 million germplasm resources (including duplicates) have been preserved [8], but the quantity of these resources led to difficulties in management, housing, and efficient studies [4]. In response to this problem, the concept of the core collection was proposed by Frankel [9], and then gradually supplemented and perfected from three aspects, including basic characteristics, construction principles and methods [10–12]. Ideally, the core collection should represent the maximum level of genetic diversity of the initial collection with the smallest number of samples. The small but highly representative sample comprising the core can facilitate in-depth genomic investigation for molecular breeding as well as other applications [13].

Since the concept of the core collection was put forward, multiple types of data have been used to construct core collection, such as phenotypic traits, DNA molecular markers. Phenotypic traits result from underlying genotypes but are affected by compound interactions between the genotype and environment. Moreover, for trees, years of continuous investigation are often needed to obtain reliable phenotypic data [14]. DNA molecular markers, which are developed based on DNA polymorphisms, represent a direct method of measuring genetic diversity and are rarely affected by environmental interactions. Thus, they are more suitable for constructing a core collection and evaluating genetic diversity than phenotypic traits [15]. For the reason that molecular technology becomes more convenient and economical, DNA molecular markers have been an advantageous and ideal tool for the establishment of core collections [16,17].

According to the definition of core collection that the least amount of germplasm represents the genetic diversity of the initial collection to the maximum extent, so the selection of core collection should avoid the germplasm with close genetic relationship. In order to do that, an appropriate and efficient sampling strategy is critical during the establishment of core collection [18]. At the same time, for the sake of verifying the representativeness of core collection, appropriate parameters are demanded to ensure genetic diversity [19]. Some sampling strategies and evaluation parameters have now been introduced to develop core collections; however, the optimal sampling strategy and evaluation parameters for each species may be different [20]. Therefore, choosing the optimal sampling strategy and a series of appropriate evaluation parameters are an important aspect of the core collection construction research.

Based on DNA molecular marker data, combining with different sampling strategies, and applying multiple evaluation parameters, core collections have been constructed for many crops, such as for *Phaseolus lunatus* [21], *Oryza sativa* [22], *Zea mays* [23,24], *Glycine max* [25], *Medicago truncatula* [26] and *Cajanus cajan* [27]. Also applied to a lot trees, such as *Cryptomeria japonica* [28], *Dalbergia Odorifera* [29], *Prunus armeniaca* [30], *Malus sieversi* [31], and *Eucalyptus cloeziana* [32]. Nevertheless, development of core collections for trees, especially those kept in the germplasm resource bank, is extremely essential. Because the preservation of this type of tree are usually maintained as living collections in the field, and this has extremely high management costs.

P. sibirica exhibits high levels of natural variation and variation introduced through selective breeding. Natural variation in *P. sibirica* results in part from its self-incompatibility, as well as human-mediated introductions into new regions followed subsequently by local adaptations. Due to the abundant variation and presumed underlying genetic diversity in *P.*

sibirica, a robust germplasm resource for the species is necessarily very large. Thus, a core collection can aid in reducing the complexity of such a large collection to the benefit of molecular breeding activities and other research. During the past 20 years, the National Forest Germplasm Resources Preservation Repository for *P. sibirica* has been operated by the Shenyang Agricultural University in Kazuo County of Liaoning Province, China. In order to better utilize this collection, accelerate the breeding process and add to its scientific and economic value, we sought in this study to establish a core collection of *P. sibirica* based on simple sequence repeat (SSR) molecular markers.

Materials and methods

Plant materials

From April 2014 to September 2019, 158 accessions of *P. sibirica* were collected from China and Russia (total 13 provenances) and were used as the initial collection from which to establish a core collection, including the characteristics of high yield, late-flowering, late-maturing, sweet flesh, double kernels, sweet almond, frost resistance, drought resistance, extreme drought resistance, pink flower, fold flower, pink anther etc. (S1 Table). The 158 accessions were preserved at the National Forest Germplasm Resources Preservation Repository for *P. sibirica* by grafting clones (Kazuo County of Liaoning Province, China; Geographical position, 119˚ 24′54E ~ 120˚ 23′24E, 40˚ 47′12N ~ 41˚ 33′53N).

Experimental site situation

The region of Kazuo country belongs to the low hilly region, the elevation of 300~400 m, the climate of which is continental monsoon. The annual average temperature is 8.7˚C, and the annual average precipitation is 491.5 mm. The average sunshine duration is 2807.8 h, and the average frost-free period is 144 d. The soil type of the experimental site is dominated by brown soil. The main woody plant resources include *Prunus sibirica*, *Prunus vulgaris*, *Prunus mandshurica*. The surrounding woody plant resources are mainly *Pinus tabuliformis*, *Robinia pseudoacacia*, *Juglans regia*, *Crataegus pinnatifida*, *Amygdalus Persica*, *Vitis vinifera*, *Salix suchowensis*, etc.

DNA extraction and primer screening

Fresh young leaves were collected from 1-year-old shoots on sun-facing sides of the 158 individual trees in mid-June and stored these in liquid nitrogen in the field and in a -80˚C freezer in the lab prior to DNA extraction. DNA was extracted with DNAsesure Plant Kit (TIANGEN).

Based on the results of Reduced-Representation Genome Sequencing (RRGS) of *P. sibirica* in 2014, 600 SSR primers were designed [33] (Beijing SBS Genetech Co., Ltd.). The primers were initially screened using three representative samples of *P. sibirica* (#354, #366, #511, respectively). Based on these preliminary screenings [34,35], 30 primer pairs (S2 Table) were selected for downstream analyses based on their polymorphism information content (*PIC*) >0.5.

PCR amplification and electrophoresis detection

The PCR amplification for SSR analysis were performed in a 20 μL PCR reaction mixture, containing 20 ng of DNA template, 0.15 μg/L of primer concentration, 2.0 mmol/L of Mg^{2+}, 1.0 U of Taq polymerase, and 0.25 mmol/L of dNTPs [33]. PCR thermocycling was performed as follows under a hot lid at 105˚C: 1) enzyme activation at 94˚C for 5 min, 2) 34 cycles of

denaturation at 94˚C for 30 s, annealing at 55˚C for 30 s, and extension at 72˚C for 30 s, 3) final extension at 72˚C for 5 min, and 4) hold at 4˚C until removed and processed [33]. The resulting PCR products were detected using non-denaturing polyacrylamide gel electrophoresis on a 12% gel. and performed electrophoresis for 90 minutes at 220 V. After rinsing, silver staining, and development, the gel was imaged in a gel imaging system (BIO-RAD, USA) (S1 Fig).

Statistical analysis of SSR data

The sizes of the amplified fragments for each locus were calculated by first comparing to a 100 bp DNA ladder (TIANGEN) in Imagelab 4.0 software and then with manual corrections based on the known size of the SSR repeat unit. Amplified fragments that differed by more than one repeat unit were identified as representing different alleles [36]. SSR dataset was converted different format in DATAtrans 2.0 for use in following analysis software. The observed alleles (N_a), the number of effective alleles (N_e), the Shannon's information index (I), and Nei's gene diversity (H) were calculated using POPGENE 32 software. Cervus 3.0 was applied to determine polymorphism information content (PIC). Cluster analysis was performed using an Unweighted Pair Group Method of Arithmetic Average (UPGMA) by NTsys V2.10e software, based on three kinds of genetic distances evaluated by using the Simple Matching genetic similarity coefficient (SM), the Jaccard genetic similarity coefficient (JD) and the Nei & Li genetic similarity coefficient (ND) [37].

Construction of the core subsets

Based on the cluster analysis, two sampling strategies (Allele preferred strategy, PS; Random strategy, RS) and three genetic distances (SM, JD, and ND) combined with stepwise clustering were used for construction of the core subsets. The (PS) strategy [31] meant that for each pair of accessions clustered in the dendrogram, the individual with more alleles was chosen for the next round of clustering. If the number of alleles was equal, the individual with more rare alleles (allele frequency $< 5\%$) was preferred to choose. If still equal, an individual was randomly chosen. The RS strategy [31] meant that the accession was randomly chosen from each pair of accessions clustered in the dendrogram to enter into the next round of clustering. When the cluster consisted of a single accession, that was the one chosen. The stepwise cluster analyses were repeated in the same way until chosen lineages were reduced to 30%, 25%, 20%, 15% and 10% of the initial collection, at which time the construction of the core subsets was complete.

Representative evaluation of the core collection

By comparing 4 genetic parameters(N_e, I, PIC and H), the most optimal ones of core subsets were chosen. A t-test for means was performed to determine if there was a difference in the three parameters between the core collection and the initial collection, and reserved collection (the other accessions except for accessions of core collection). The representativeness of the core collection was also validated by principal coordinates analysis (PCOA) using NTSYS V2.10e, with the distribution of the initial collection and the core collection being plotted by the first principal component score and the second principal component score.

Results

Genetic diversity of *P. sibirica*

The genetic diversity of the initial collection was analyzed based on 30 pairs of SSR primers (S3 Table). The results of partial primer polyacrylamide were shown in S3 Table. The average of 20

alleles for each pair of primers was detected in each accession, and the range was 9 to 36. The analysis of genetic diversity of the 158 accessions indicated that the average of the observed number of alleles (N_a), the effective number of alleles (N_e), Shannon's information index (I), polymorphic information content (PIC), and Nei's gene diversity (H) were 22, 8.194, 2.328, 0.847, 0.854, respectively. These parameters clearly indicated that the 158 *P*. *sibirica* germplasms had a high genetic diversity at the molecular level.

Selection of core subsets

Based on SSR data and UPGMA clustering results, 30 core subsets were constructed using stepwise clustering with two sampling strategies, three genetic distances and five sample sizes (S4 Table). First, the values of N_e, I, PIC and H of two sampling strategies were compared, to determine which sampling strategy could be used to construct the core collection (S4 Table). When using SM genetic distance to cluster, the indices of N_e, I, PIC and H of PS strategy were almost larger than the RS, except the values at 30% sampling size. When using JD and ND to cluster, the values of N_e, I, PIC and H of PS were all higher than the RS strategy. In addition, for PS strategy, the standard deviation (STDV) and coefficient of variability (CV) of N_e were 0.370 and 0.043, respectively, and the STDV and CV of I were 0.415 and 0.054, respectively (S4 Table). All of these indices were less than those obtained using RS strategies. Therefore, PS strategies was more appropriate than RS strategies as a sampling strategy to construct a core collection of *P*. *sibirica*.

Further, the values of N_e, I, PIC and H of the core subsets were also compared constructed using seven different sample sizes (10%, 15%, 20%, 25%, 30%) generated according to the PS strategy among three genetic distances (Fig 1). When using SM, JD and PS strategy, the values of N_e, I, PIC and H gradually increased and then decreased with the increase of sampling size, reaching the peak values at 25% sampling size, respectively (Fig 1A–1D). Similarly, using ND and PS strategy, the values of N_e, I, and PIC also reached their maximal value at the 25% sample size (Fig 1A–1C); although H did not significantly change, it also had the maximal value at the 25% sample size(Fig 1D). Thus, the 25% sample size was the most suitable for construction of the core collection using SM, JD and ND, respectively.

Finally, the values of N_e, I, PIC and H of core subsets using SM with 25% sampling size, JD with sampling size and ND with 25% sampling size were compared, to determine which genetic distance was the most suitable for establish the core collection (Fig 2A–2D). Obviously, the values of Ne, I, PIC and H of ND-25% were all higher than SM-25% and JD-25%. Therefore, the core subset constructed using the PS strategy with 25% sample size and ND genetic distance had the best representativeness of the initial collection in the 30 core subsets and could be used as the core collection of *P*. *Sibirica*.

Evaluation and confirmation of core collection

Results of T-test comparing three genetic diversity parameters between the core collection and the initial collection, indicated that the core collection was no significantly different from the initial collection (Table 1). Similarly, the indices except N_e of the core collection were no significantly difference with reserved collection (Table 1). The percentages of retention of N_e, I, PIC and H of the core collection were 113%, 102%, 102% and 103%, which were higher than those of the reserved collection (91%, 96%, 98%, and 99% respectively, Table 2). These results showed that the core collection has higher polymorphism and genetic diversity than the initial collection and reserved collection.

The geographical distribution of core collection was further analyzed. Each of the 13 provenances all had germplasm selected as core collection. There were 1, 2, 3, 2, 1, 1, 7, 5, 5, 4, 5, 3, 1

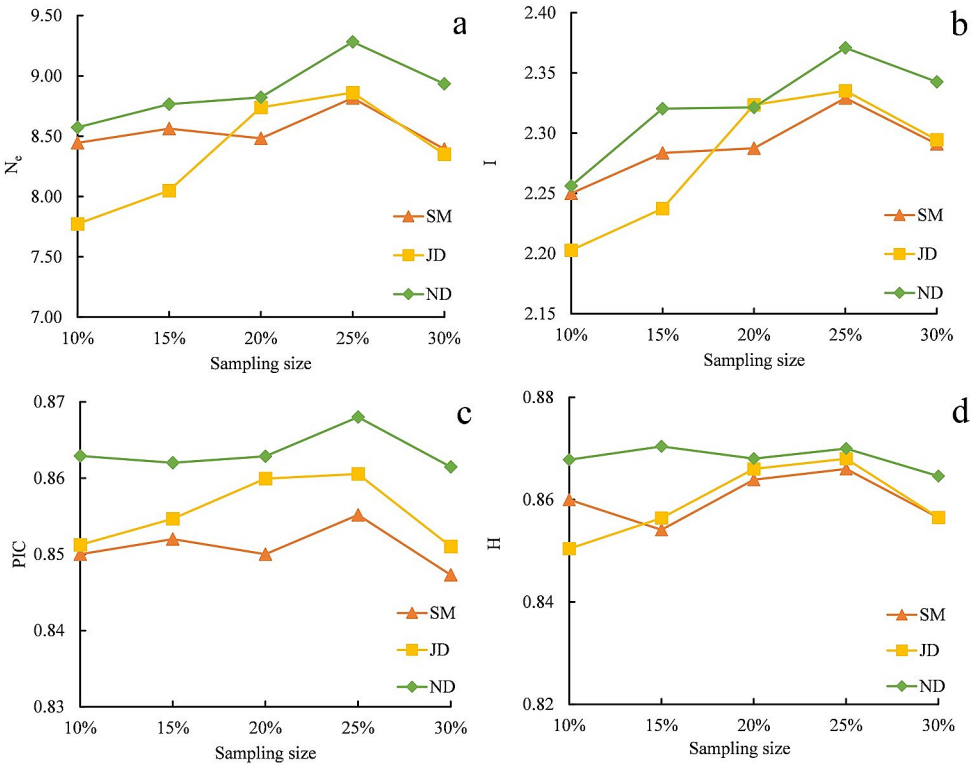

Fig 1. Values of N_e, I, PIC and H of core subset based on different sample size and different genetic distance with allele preferred strategy. ND: Genetic distance using Nei & Li genetic similarity coefficient, SM: Genetic distance using simple matching coefficient, JD: Genetic distance using Jaccard genetic similarity coefficient, N_E: Number of effective alleles, I: Shannon's information index, PIC: Polymorphic information content, H: Nei's gene diversity.

https://doi.org/10.1371/journal.pone.0260097.g001

accessions were selected into the core collection, respectively, from BJ, HLJ, HLP, HWC, HZL, IAH, IZLT, JL, LBP, LCY, LKZ, R, SY (Fig 3). The results of PCOA showed that the selected samples within the proposed core collection were distributed in a scattered pattern among the 158 total samples (Fig 4). The Fig 4 showed that the distribution pattern of the core collection was very similar to that of the initial collection, and more peripheral individuals were selected, providing further evidence of the representativeness of the proposed core. A total of 158 accessions contained 14 superior and variant types. After selective extraction, the 40 core collections still high yield, double kernel, sweet kernel, cold resistance, late flower, sweet meat, drought tolerance, extreme drought tolerance, late maturity, late-flowering (S1 Table). These results indicated the representativeness of the core collection constructed by allele preferred strategy combined with 25% sampling size. Thus, priority should be given to research and utilization of accessions represented within this newly established core collection of *P. sibirica*.

Discussion

Molecular data of constructing a core collection

At present, phenotypic traits and molecular markers are often used to construct core collections [5]. Phenotypic traits result from underlying genotypes but are affected by compound interactions between the genotype and environment. Therefore, phenotypic traits are not always good proxies for genetic diversity and cannot fully reflect the genetic diversity of an entire germplasm collection. Moreover, for trees, years of continuous investigation are often

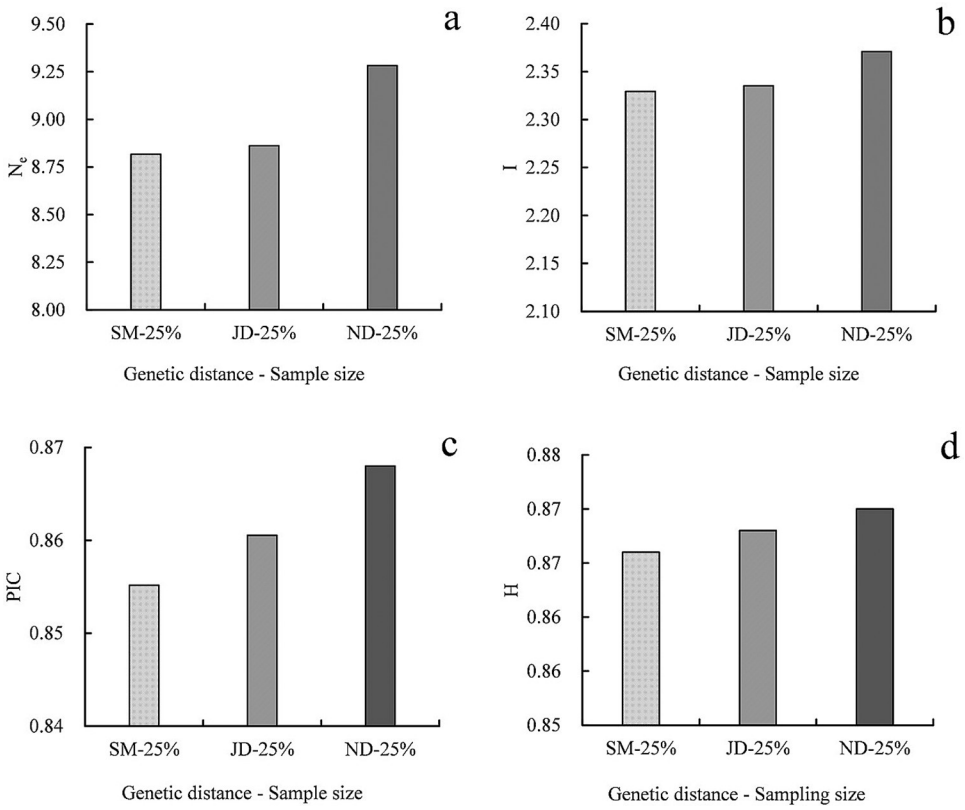

Fig 2. Values of Ne, I, PIC and H of core subsets based on the allele preferred strategy and different genetic distances with a given sample size. N_e: Number of effective alleles, I: Shannon's information index, PIC: Polymorphic information content, H: Nei's gene diversity, ND: Genetic distance using Nei & Li genetic similarity coefficient, SM: Genetic distance using simple matching coefficient, JD: Genetic distance using Jaccard genetic similarity coefficient.

https://doi.org/10.1371/journal.pone.0260097.g002

needed to obtain reliable phenotypic data [14]. Molecular markers, which are developed based on DNA polymorphisms, represent a direct method of measuring genetic diversity and are rarely affected by environmental interactions. Therefore, they are more suitable for

Table 1. T-test results of the optimal core collection and initial collection, core collection and reserved collection.

Parameter	Germplasm	Mean	STDV	Pair mean	Pair deviation	T value	Sig.
N_e	Initial collection	8.194	3.099	-1.088	0.895	-1.215	0.229
	Reserved collection	7.489	2.818	-1.793	0.863	-2.077	0.042*
	Core collection	9.282	3.797				
I	Initial collection	2.328	0.403	-0.043	0.104	-0.412	0.682
	Reserved collection	2.240	0.396	-0.131	0.103	-1.270	0.209
	Core collection	2.371	0.400				
PIC	Initial collection	0.847	0.086	-0.021	0.020	-1.013	0.315
	Reserved collection	0.831	0.095	-0.036	0.217	-1.674	0.099
	Core collection	0.868	0.071				
H	Initial collection	0.854	0.076	-0.016	0.018	-0.906	0.369
	Reserved collection	0.840	0.095	-0.030	0.019	-1.564	0.123
	Core collection	0.870	0.063				

N_e: Number of effective alleles, I: Shannon's information index, PIC: Polymorphic information content, H: Nei's gene diversity.

https://doi.org/10.1371/journal.pone.0260097.t001

Table 2. Contradistinction of the genetic diversity between initial collection, core collection and reserved collection.

Germplasm	Numbers of germplasm	N_a	N_e	I	PIC	H
Initial collection	158	22	8.194	2.328	0.847	0.854
Core collection	40	16	9.282	2.371	0.868	0.870
Percentage of retention	25.3%	73%	113%	102%	102%	103%
Reserved collection	118	18	7.489	2.240	0.831	0.840
Percentage of retention	74.7%	82%	91%	96%	98%	99%

N_e: Number of effective alleles, I: Shannon's information index, PIC: Polymorphic information content. H: Nei's gene diversity.

https://doi.org/10.1371/journal.pone.0260097.t002

constructing a core collection and evaluating genetic diversity than phenotypic traits [15]. Molecular markers are divided into dominant and codominant types. SSRs are a codominant type of marker and consequently can provide richer allelic information than dominate markers [38]. Therefore, 30 SSR primers with high polymorphism were used to obtain 40 accessions of *P. sibirica* as a core collection to represent the initial collection. The successful establishment of core collection was of great significance to the further research and utilization of *P. Sibirica*. germplasm resources.

Methods for constructing the core collection

Sampling strategy is critical to construction of a core collection. Since the concept of core collection was proposed, many scholars have given different suggestions on its construction methods, such as the method of maximizing the number of alleles [39], and the clustering method based on genetic distance [40], rare allele priority method (PS) [31], etc. So far, rare allele priority method is one of the most common methods. It has been used in some forest tree species, for instance *Ficus carica L.* [41], *Castanea mollissima* [42], *Armeniaca vulgaris* Lam [43] and so on. Ideally, a core collection should represent the genetic diversity of the

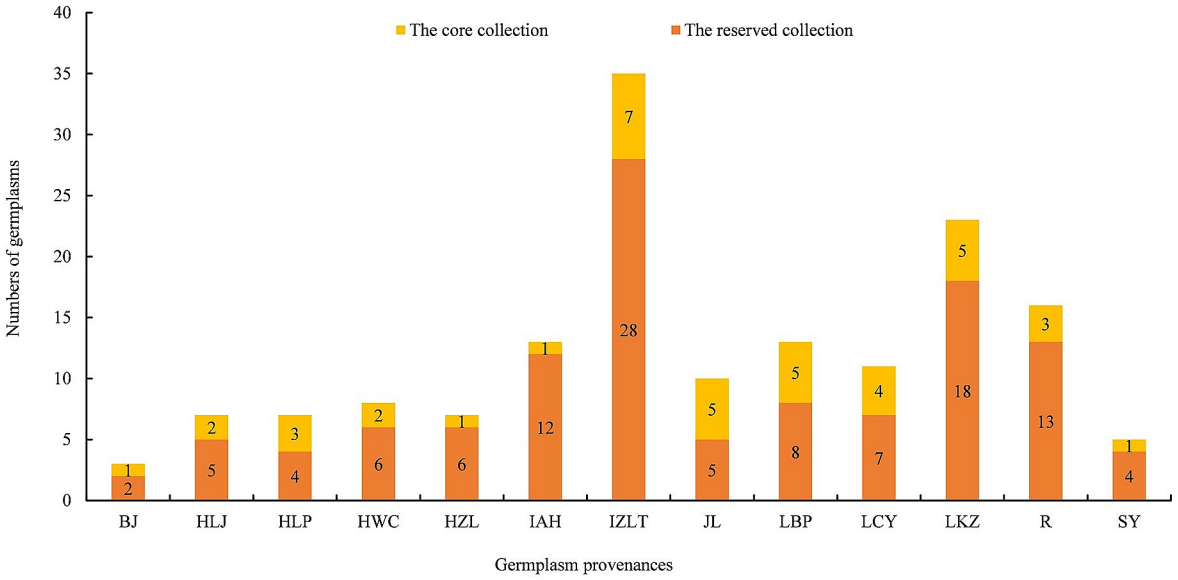

Fig 3. Numbers of core collection and reserved collection of each provenance. The numbers of core collection (yellow) plus the numbers of reserved collection (orange) were equal to the numbers of initial collection. BJ: Beijing; HLJ: Heilongjiang; HLP: Luanping, Hebei; HWC: Weichang, Hebei; HZL: Zhuolu, Hebei; IAH: Aohan, Inner Mongolia; IZLT: Zhalantun, Inner Mongolia; JL: Jilin; LBP: Beipiao, Liaoning; LCY: Chaoyang, Liaoning; LKZ: Kazuo, Liaoning; R: Russia; SY: Yuxian, Shanxi.

https://doi.org/10.1371/journal.pone.0260097.g003

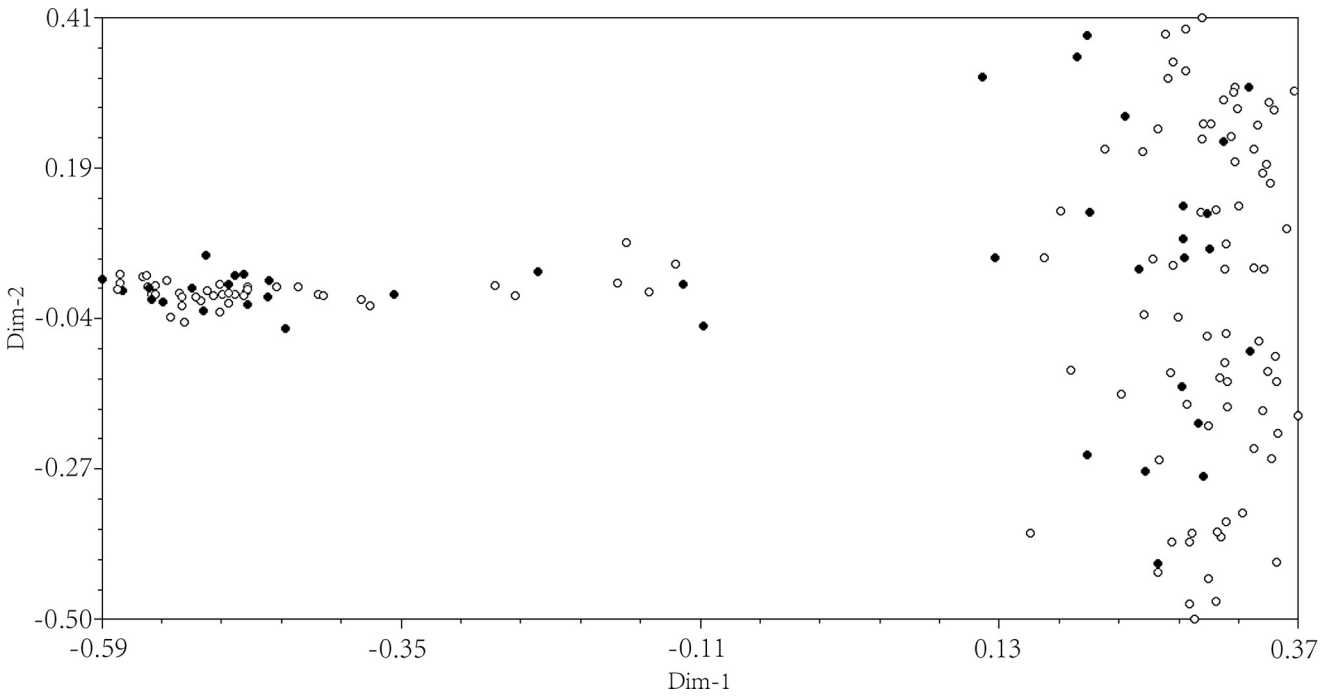

Fig 4. Principal coordinate plot of the core collection constructed using the preferred sampling strategy with 25% sample size and Nei & Li genetic distance and initial collection. The 118 accessions of the reserved collection were represented by open circles and the 40 accessions of the core collection were represented by black circles. All circles represented the 158 of the initial collection. Dim-1 and Dim-2 represented the first principal component score and the second principal component score of 158 accessions.

https://doi.org/10.1371/journal.pone.0260097.g004

entire germplasm resource as comprehensively as possible, and the genetic diversity of the entire germplasm resource depends largely on the number of alleles and allele frequencies at all loci [44]. Therefore, rare alleles (allele frequency < 5%) should preferentially be selected when selecting among germplasm to develop a core collection. As a result, it is widely acknowledged that PS strategy is likely better than a RS strategy for most species [31,43,45]. Consistent with previous results, in this study, we found that PS yielded higher values for N_e, I, PIC and H than the RS in all candidate core collections. Therefore, PS is a better sampling strategy for constructing the core collection of *P. sibirica*.

It is important to confirm a reasonable sampling percentage during core collection construction. Brown proposed that an ideal core collection size should be about 10% of the total collection, which maintained over 70% of the alleles in the whole collection with 95% certainty [9]. Yoneazwa et al. suggested that the optimal proportion was 20–30% [46]. However, The sample percentage of a typical core collection is generally between 5% and 40% of the germplasm resource [4,47]. Generally, the appropriate sampling percentage should be determined according to the characteristics of different germplasm populations; the initial germplasm population is small, it is most suitable to choose a large sampling percentage, while the initial germplasm population is large, it is most suitable to choose a larger sampling percentage [29,37]. Gomes et al. analyzed the genetic diversity of 153 lima bean breeding lines and obtained a core collection of 34 lines, accounting for 22% of the initial collection [21]. Hu et al. analyzed the genetic diversity of 612 accessions of *Cucumis melo* L. and obtained a core collection of 118 accessions, accounting for 19.4% of the initial collection [48]. Mongkolporn et al. analyzed the genetic diversity of 230 *Capsicum* spp. and obtained a core collection of 28 accessions, accounting for 12% of the initial collection [49]. These results indicated that the

sampling percentage should depend on the quantity and genetic diversity of the germplasm population. According to the germplasm population size in this study, five sampling percentages (10%, 15%, 20%, 25%, 30%) were set to establish the core collection. By comparing values of N_e, I, PIC and H among the core collections constructed with the 5 sample sizes using the PS strategy and the same genetic distance, the optimal proportion was found to be 25% for SM, JD and ND, respectively.

Genetic distance is a measure of the genetic difference between species in a population. The genetic distance measures, SM, JD, and ND, are commonly used the clustering of molecular markers such as into dendrograms and are known to produce different results [31]. So, genetic distance will affect the results of clustering and the construction of a core collection [47]. In this study, the Ne, I, PIC and H values of the core collection constructed with 25% sample size with ND genetic distance were higher than those of other methods. Therefore, the core collection constructed by the PS strategy using 25% sample size and ND genetic distance was considered to be the core collection of *P. sibirica*.

Evaluation of the proposed core collection

Representativeness is the most important property of core collection. Commonly used core collection genetic diversity evaluation parameters are: Number of alleles, Shannon diversity index and Nei's gene diversity index. The number of alleles was considered the most relevant indicator [50,51]. Wang et al. considered that Shannon diversity index, Polymorphic information content and Simpson diversity index were important parameters for evaluating the representativeness of core collection when studying the evaluation parameters of rice core germplasm [47]. Some scholars have also used parameters such as the percentage of polymorphic loci, the number of observed alleles, and the number of effective alleles to evaluate the genetic diversity of core collections [47,52]. Based on previous research results, this study selected 4 parameters (N_a, N_e, I, PIC, H) and their retention ratios, combined with the t-test and principal coordinate analysis (PCO), to verify and confirm the optimal core collection, and the effect is better. The retention ratio was the percentage of each genetic parameter of the core collection to each genetic parameter of the initial collection [53]. In this study, the retention ratios of 4 parameters (Ne, I, PIC, H) were greater than 100%, which was mainly caused by changes in the number of samples and allele frequencies in the population. These genetic parameters were estimated by different methods based on the allele frequency, and were used to measure the genetic redundancy in the diversity of different populations (initial collection, core collection, and reserved collection). The construction process of core collection is a process of reducing the frequency of alleles and increasing the proportion of rare alleles. In the process of removing genetic redundancy, the irregular increase and decrease of the frequency of each allele can easily lead to the retention rate of the corresponding genetic diversity parameter being greater than 100%. This phenomenon was common in the core collections of *Armeniaca vulgaris* Lam. [43], and *Eucommia ulmoides* [53], exist. According to the allele retention ratio must be greater than 70%, the larger the other genetic parameters, the better the evaluation criteria [27,54]. In the case that the retention ratios of N_a were 73%, the core collection constructed in this study has better genetic parameters and meet the requirements of the core collection, indicating that the effect of retaining the genetic diversity of the initial collection was better.

Characteristics of germplasms

Siberian apricot breeding process hinge on the abundant germplasm resources, and understanding the characteristics of resources is the prerequisite for effective utilization of apricot

germplasm resources. In the past 20 years, 158 Siberian apricot germplasm resources collected by our research group were characterized and evaluated. The germplasm resources of Siberian apricot were rich in characteristics, including high yield, frost resistance, drought resistance, bent branch, late-maturing, late-flowering, fold flower etc. Among those characteristics, 70.25% germplasms had the potential of high yield. Approximately 8% of the germplasm has the characteristics of late flowering or barren tolerance, which makes it adapt to a more complex ecological environment. A small amount of Siberian apricot also has attractive characteristics on branches, petals, fruits, and almond and germplasm with these characteristics could be very useful in food, ornamental and other aspects. All in all, Siberian apricot variation is abundant. We used these characteristics information to evaluate the core collection based on SSR, and the results showed that the core collection contains most of the characteristic types.

Conclusions

In this study, we tested combinations of sampling strategies, measures of genetic diversity, and sample reduction techniques to find the one most suited to this species using SSR markers data. Based on these tests, we established a core collection based on an allele preferred sampling strategy, the Nei & Li genetic distance, and 25% sampling from the complete resource. The new core collection comprises of 40 *P. sibirica* from 13 geographical regions and could represent the genetic diversity of the complete germplasm collection. In the next step, we will focus on the investigation of phenotypic traits of core collection to provide more valuable information for the development and utilization of germplasm resources of Siberian apricot, which will lead to better utilization of germplasm for Siberian apricot breeding programs. In addition, the core collection is further modification by continuously adding new germplasm resources, which will give interesting insights about the representation of unknown Siberian apricot diversity.

Supporting information

S1 Fig. The polyacrylamide gel electrophoresis of PCR products amplified by SSR primers.
(PDF)

S1 Table. The *Prunus sibirica* germplasm resources for test materials.
(DOCX)

S2 Table. The information of 30 SSR markers.
(DOCX)

S3 Table. Diversity index at 30 SSR loci in *Prunus sibirica*.
(DOCX)

S4 Table. The genetic diversity of the initial collection and 30 core subsets.
(DOCX)

Acknowledgments

The authors would like to thank TopEdit (www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript.

Author Contributions

Conceptualization: Yongqiang Sun, Shengjun Dong.

Formal analysis: Yongqiang Sun, Jianhua Chen, Jingjing Pan.

Funding acquisition: Shengjun Dong.

Methodology: Yongqiang Sun, Jian Zhang.

Validation: Yongqiang Sun, Jianhua Chen.

Writing – original draft: Yongqiang Sun.

Writing – review & editing: Shengjun Dong, Quangang Liu.

References

1. Niu J, Bi QX, Deng SY, Chen HP, Yu HY, Wang LB, et al. Identification of *AUXIN RESPONSE FACTOR* gene family from *Prunus sibirica* and its expression analysis during mesocarp and kernel development. BMC Plant Biology. 2018; 18(1):21. https://doi.org/10.1186/s12870-017-1220-2 PMID: 29368590
2. Bazha SN, Baskhaeva TG, Danzhalova EV, Drobyshev YI, Ivanov LA, Ivanova LA, et al. Ecological and Biological Features of the Distribution of the Siberian Apricot (*Prunus sibirica* L.) in the Southern Part of the Selenga River Basin. Arid Ecosystems. 2020; 10(4):284–292. http://doi.org/doi:10.1134/S2079096120040022.
3. Li M, Zhao Z, Miao X, Zhou J. Genetic Diversity and Population Structure of Siberian apricot (*Prunus sibirica* L.) in China. International Journal Molecular Sciences. 2014; 15(1):377–400. https://doi.org/10.1371/journal.pone.0087381 PMID: 24516551
4. Escribano P, Viruel MA, Hormaza JI. 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 underutilized subtropical fruit tree species. Annals of Applied Biology. 2008; 153(1):25–32. https://doi.org/10.1111/j.1744-7348.2008.00232.x.
5. Duan H, Cao S, Zheng H, Hu D, Lin J, Cui B, et al. Genetic Characterization of Chinese fir from Six Provinces in Southern China and Construction of a Core Collection. Scientific Reports. 2017; 7(1):13814. https://doi.org/10.1038/s41598-017-13219-0 PMID: 29062029
6. Gecer MK, Kan T, Gundogdu M, Ercisli S, Ilhan G, Sagbas HI. Physicochemical characteristics of wild and cultivated apricots (*Prunus armeniaca* L.) from Aras valley in Turkey. Genetic Resources and Crop Evolution, 2020; 67(2):935–945. http://doi.org/10.1007/s10722-020-00893-9.
7. Karatas N, Sengul M. Some important physicochemical and bioactive characteristics of the main apricot cultivars from Turkey. Turkish Journal of Agriculture and Forestry, 2020; 44(6):651–661. http://doi.org/10.3906/tar-2002-95.
8. Miao LM, Wang SY, Zou MH, Li JB, Kong LJ, Yu XJ. Review of the studies on core collection for horticultural crops. Journal of Plant Genetic Resources. 2016; 17(5):791–800. (in Chinese) http://doi.org/10.13430/j.cnki.jpgr.2016.05.001.
9. Frankel OH. Genetic Perspectives of Germplasm Conservation. Cambridge: Cambridge University Press; 1984.
10. Brown AHD. Core collection: a practical approach to genetic resources management. Genome.1989; 31(2):818–824. https://doi.org/10.1139/g89-144.
11. Basigalup DH, Barnes DK, Stucker RE. Development of a core collection for perennial Medicago plant introductions. Crop Science. 1995; 35(4):1163–1168. http://doi.org/10.2135/cropsci1995.0011183X003500040042x.
12. Casler MD, Santen EV. Patterns of variation in a Collection of meadow fescue accessions. Crop Science. 2000; 40(1):248–255. http://doi.org/10.2135/cropsci2000.401248x.
13. Belaj A, Dominguez-Garcia M D, Atienza SG, Urdiroz NM, Rosa RDL, Satovic Z, et al. Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. Tree Genetics and Genomes. 2012; 8(2):365–378. http://doi.org/10.1007/s11295-011-0447-6.
14. Wang HX, Zhao SG, Gao Y, Xuan LC, Zhang ZH. A construction of the core-collection of *Juglans regia* L. based on AFLP molecular markers. Scientia Agricultura Sinica. 2013; 46(23):4985–4995. (in Chinese) http://doi.org/10.3864/j.issn.0578-1752.2013.23.015.
15. Liu XL, Liu HB, Ma L, Li XJ, Xu CH, Su HS, et al. Construction of sugarcane hybrids core collection by using stepwise clustering sampling approach with molecular marker data. Acta Agronomica Sinica. 2014; 40(11):1885–1894. (in Chinese) http://doi.org/10.3724/SP.J.1006.2014.01885.

16. Li Y, Li YH, Yang QW, Zhang JP, Zhang JM, Qiu LJ, et al. Genomics-based Crop Germplasm Research: Advances and Perspectives. Scientia Agricultura Sinica. 2015; 48(7):3333–3353. (in Chinese) http://doi.org/10.3864/j.issn.0578-1752.2015.17.003.

17. Van Treuren R, Tchoudinova I, Van Soest LJM, Van Hintum ThJL. Marker-assisted acquisition and core collection formation: a case study in barley using AFLPs and pedigree data. Genetic Resources and Crop Evolution. 2006; 53(1):43–52. http://doi.org/10.1007/s10722-004-0585-x.

18. Odong TL, Jansen J, Van Eeuwijk FA, Van Hintum TJL. Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theoretical and Applied Genetics. 2013; 126(2):289–305. https://doi.org/10.1007/s00122-012-1971-y PMID: 22983567

19. Wang JC, Hu J, Zhang CF, Zhang S. Assessment on evaluating parameters of rice core collections constructed by genotypic values and molecular marker information. Rice Science. 2007a; 14(2):101–110. http://doi.org/10.1016/S1672-6308(07)60015-8.

20. Van Hintum ThJL, Brown AHD, Spillane C, Hodgkin T. Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy. 2000. http://doi.org/10.1016/S0304-4238(96)00927-2.

21. Gomes RLF, Costa MF, Alvespereira A, Bajay MM, Zucchi MI. A lima bean core collection based on molecular markers. Scientia Agricola. 2020; 77(2):1–8. http://doi.org/10.1590/1678-992x-2018-0140.

22. Zhang HL, Zhang DL, Wang MX, Sun JL, Qi YW, Li JJ, et al. A core collection and mini core collection of *Oryza sativa* L. in China. Theoretical and Applied Genetics. 2011; 122(1):49–61. https://doi.org/10.1007/s00122-010-1421-7 PMID: 20717799

23. Li Y, Shi YS, Cao YS, Wang TY. Establishment of a Core Collection for Maize Germplasm Preserved in Chinese National Gene Bank using Geographic Distribution and Characterization Data. Genetic Resource and Crop Evolution. 2004; 51(8):845–852. http://doi.org/10.1007/s10722-005-8313-8.

24. Coimbra RR, Miranda GV, Cruz CD, Silva DJ, Vilela RA. Development of a Brazilian maize core collection. Genetics and Molecular Biology. 2009; 32(3):538–545. https://doi.org/10.1590/S1415-47572009005000059 PMID: 21637517

25. Wang LX, Guan Y, Guan RX, Li YH, Ma YS, Dong ZM, et al. Establishment of Chinese soybean (*Glycine max*) core collections with agronomic traits and SSR markers. Euphytica. 2006; 151(2):215–223. http://doi.org/10.1007/s10681-006-9142-3.

26. Ronfort J, Bataillon T, Santoni S, Delatande M, David JL, Prosperi JM. Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula*. BMC Plant Biology. 2006; 6(1):28. http://doi.org/10.1186/1471-2229-6-28.

27. Reddy LJ, Upadhyaya HD, Gowda CLL, Singh S. Development of core collection in pigeonpea [*Cajanus cajan* (L.) Millspaugh] using geographic and qualitative morphological descriptors. Genetic Resource and Crop Evolution. 2005; 52(8):1049–1056. http://doi.org/10.1007/s10722-004-6152-7.

28. Miyamoto N, Ono M, Watanabe A. Construction of a core collection and evaluation of genetic resources for *Cryptomeria japonica* (Japanese cedar). Journal of Forest Research. 2015; 20(1):186–196. http://doi.org/10.1007/s10310-014-0460-3.

29. Liu FM, Zhang NN, Liu XJ, Yang ZJ, Jia HY, Xu DP. Genetic diversity and population structure analysis of *Dalbergia Odorifera* germplasm and development of a core collection using microsatellite markers. Genes. 2019; 10(4): 281. https://doi.org/10.3390/genes10040281 PMID: 30959931

30. Wang YZ, Zhang JH, Sun HY, Ning N. Construction and evaluation of a primary core collection of apricot germplasm in China. Scientia Horticulturae. 2011; 128(3):311–319. http://doi.org/10.1016/j.scienta.2011.01.025.

31. Zhang CY. Chen XS, Zhang YM, Yuan ZH, Liu ZC, Wang YL, et al. A method for constructing core collection of *Malus sieversii* using molecular markers. Agricultural sciences in China. 2009; 8(3):267–284. http://doi.org/10.1016/S1671-2927(08)60210-2.

32. Lv JB, Li CR, Zhou CP, Chen JB, Li FG, Wang QG, et al. Genetic diversity analysis of a breeding population of Eucalyptus cloeziana F. Muell. (Myrtaceae) and extraction of a core germplasm collection using microsatellite markers. Industrial Crops and Products, 2020, 145:112157. http://doi.org/10.1016/j.indcrop.2020.112157.

33. Chen JH, Dong SJ, Zhang X, Wu YL, Zhang HK, Sun YQ, et al. Genetic diversity of *Prunus sibirica* L. superior accessions based on the SSR markers developed using restriction-site associated DNA sequencing. Genetic resources and crop evolution, 2021; 68(2): 615–628. https://doi.org/10.1007/s10722-020-01011-5.

34. Zhang HK. Study on germplasm resources diversity in Armeniaca mandshurica. M.A. Thesis, Shenyang Agricultural University. 2017.

35. Lu CY. Genetic diversity based on SSR and its association analysis with phenotypic traits in Armeniaca vulgaris var. ansu. M.A. Thesis, Shenyang Agricultural University. 2018.

36. Wen ZX, Zhao TJ, Zheng YZ, Liu SH, Wang CE, Wang F, et al. Association analysis of agronomic and quality traits with SSR markers in Glycine max and Glycine soja in China: I. population structure and associated markers. Acta Agronomica Sinica. 2008; 34(7):1169–1178. (in Chinese) http://doi.org/10.1016/S1875-2780(08)90000-6.

37. Xu Y, Chen CS, Ji DH, Xu K, Xie XX, Xie CT. Developing a core collection of *Pyropia haitanensis* using simple sequence repeat markers. Aquaculture. 2016; 452:351–356. http://doi.org/10.1016/j.aquaculture.2015.11.016.

38. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics. 2004; 5(6):435–45. https://doi.org/10.1038/nrg1348 PMID: 15153996

39. Schoen DJ, Brown AHD. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proceedings of the National Academy of Sciences. 1993; 90(22):10623–10627. https://doi.org/10.1073/pnas.90.22.10623 PMID: 8248153

40. Dervishi A, Jakše J, Ismaili H, Javornik B, Štajner N. Genetic Structure and Core Collection of Olive Germplasm from Albania Revealed by Microsatellite Markers. Genes. 2021; 12(2): 256. https://doi.org/10.3390/genes12020256 PMID: 33578843

41. Balas FC, Osuna MD, Domínguez G, Pérez-Gragera F, López-Corrales M. Ex situ conservation of underutilised fruit tree species: establishment of a core collection for *Ficus carica* L. using microsatellite markers (SSRs). Tree Genetics and Genomes. 2014;703–710. https://doi.org/10.1007/s11295-014-0715-3.

42. Nie XH, Wang ZH, Liu NW, Song L, Cao QQ. Fingerprinting 146 Chinese chestnut (*Castanea mollissima* Blume) accessions and selecting a core collection using SSR markers. Journal of Integrative Agriculture, 2021, 20(5):1277–1286. https://doi.org/10.1016/S2095-3119(20)63400-1.

43. Liu J, Liao K, Zhao SR, Cao Q, Sun Q, Liu H. Core-germplasm construction of apricot collections in south of Xinjiang by ISSR molecular markers. Journal of Fruit Science. 2015; 32(5):374–384. (in Chinese) http://doi.org/10.13925/j.cnki.gsxb.20140447.

44. Guo D L, Liu CH, Zhang JY, Zhang GH. Construction of grape core collection. Scientia Agricultura Sinica. 2012; 45(6):1135–1143. (in Chinese) http://doi.org/10.3864/j.issn.0578-1752.2012.06.011.

45. Duan F, Zhang H, Li S, Tian ZQ, Gan XH. Core collection construction of endangered plant *Tetracentron sinense* based on ISSR molecular markers. Subtropical Plant Science. 2018; 47(2):101–106. http://doi.org/10.3969/j.issn.1009-7791.2018.02.001.

46. Yoneazwa K, Nomura T, Morish H. Sampling strategies for use in stratified germplasm collection. In: Hodgkin T, Brown AHD, Van Hintum THL, editors. Core Collection of Plant Genetic Resources. Chichester: John Wily and Sons.1995; pp.35–53.

47. Wang JC, Hu J, Xu HM, Zhang S. A strategy on constructing core collections by least distance stepwise sampling. Theoretical and Applied Genetics. 2007b; 115(1):1–8. https://doi.org/10.1007/s00122-007-0533-1 PMID: 17404701

48. Hu JB, Wang PQ, Su Y, Wang RJ, Li Q, Sun KL. Microsatellite Diversity, Population Structure, and Core Collection Formation in Melon Germplasm. Plant Molecular Biology Reporter. 2014; 33(3):439–447. http://doi.org/10.1007/s11105-014-0757-6.

49. Mongkolporn O, Hanyong S, Chunwongse J, Wasee S. Establishment of a core collection of Chilli germplasm using microsatellite analysis. Plant Genetic Resources. 2015; 13(2):104–110. http://doi.org/10.1017/S1479262114000768.

50. Mousadik AE, Petit R J. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. Theoretical and Applied Genetics. 1996; 92(7):832–839. https://doi.org/10.1007/BF00221895 PMID: 24166548

51. Petit RJ, Mousadik AE, Pons O. Identifying populations for conservation on the basis of genetic markers. Conservation Biology. 1998; 12: 844–855. http://doi.org/10.1111/j.1523-1739.1998.96489.x.

52. Wang X, Cao Z, Gao C, Li K. Strategy for the construction of a core collection for *Pinus yunnanensis* Franch. to optimize timber based on combined phenotype and molecular marker data. Genetic Resources and Crop Evolution. 2021; 1–22. https://doi.org/10.1007/s10722-021-01182-9.

53. Li HG, Xu JH, Du HY, Wuyun TN, Liu PF, Du QX. Preliminary Construction of Core Collection of *Eucommia ulmoides* Based on Allele Number Maximization Strategy. Scientia Silvae Sinicae. 2018; 54(2):42–51. (in Chinese) http://doi.org/10.11707/j.1001-7488.20180205.

54. Mckhann HI, Camilleri C, Bérard A, Bataillon T,David JL, Reboud X, et al. Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. The Plant Journal. 2004; 38(1):193–202. https://doi.org/10.1111/j.1365-313X.2004.02034.x PMID: 15053772