Journal of Tropical Forest Science 35(4): 476–488 (2023)
https://doi.org/10.26525/jtfs2023.35.4.476
ISSN: 0128-1283, eISSN: 2521-9847

CREATING A CORE COLLECTION OF SUPERIOR *ALNUS NEPALENSIS* TREES BASED ON PHENOTYPE AND MOLECULAR MARKER DATA

Wang XL, Zou GQ & Cao ZL*

Key Laboratory for Forest Resources Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming 650224, China

**fjcaozilin@sina.com*

Submitted January 2023; accepted March 2023

Towards the improved preservation and management of high-quality germplasm resources for *Alnus nepalensis*, a total of 84 samples from 28 superior *A. nepalensis* trees were used as an original germplasm collection. An optimal strategy for the creation of a core collection of superior *A. nepalensis* trees was devised by comparing the validity of 18 germplasm subsets. Each of these germplasm subsets was configured using unweighted average clustering and improved minimum distance stepwise sampling methods that integrated six sampling proportions, three data types, and three genetic distance methods. The germplasm subsets for the original germplasm collection were evaluated by quantitative and qualitative trait independence tests according to the data types. The results revealed that the developed germplasm subsets based on phenotype data and Euclidean genetic distances at 55, 45, 35 and 25% sampling proportions; molecular marker data and Nei's genetic distances at 55, 45, 35, 25 and 15% sampling proportions; and combined phenotype and molecular marker data and mixed genetic distances at 55, 45, 35 and 25% sampling proportions could represent the original germplasm collection. The strategy of phenotype values and Euclidean genetic distance were more appropriate than integrating the phenotype and molecular markers data and mixed genetic distance for the creation of a core collection for superior *A. nepalensis* trees.Considering the effectiveness, practicability, and cost of the development of a core collection for superior *A. nepalensis* trees, an optimal strategy included an unweighted average technique, improved minimum distance stepwise sampling method, SRAP markers data, Nei's genetic distance, and a 15% sampling proportion.

Keywords: Quantitative and qualitative trait independence tests, sampling proportion, data type, genetic distance methods

INTRODUCTION

Many crop germplasm banks have been developed to protect the increasing loss of genetic diversity of germplasm resources (Arriel et al. 2023). Germplasm banks have become steadily larger through the continuous collection of germplasm resources. Therefore, the management burden of germplasm banks has increased but the genetic variation of germplasm resources has not. Frankel (1984) proposed the concept of core collections, which can mediate the direction of plant germplasm resource conservation. To date, more than 100 core collections for 80 plant species have been developed worldwide. However, research into core collections for perennial woody plants is very poor. For trees, the research into core collections for economic tree species (e.g. *Litchi chinensis.*, *Juglans regia, Ziziphus jujuba, Malus sieversii, Canarium album, Vitis vinifera*) are greater than those for timber tree species (*Pinus yunnanensis, Populus deltoides, Cinnamomum camphora, Pinus massoniana*, etc.) (Li et al. 2020, Zhang et al. 2018, Guardo et al. 2019, Wang et al. 2021). Thus far, research into a core collection for *Alnus nepalensis* has not been conducted.

In 2010, a provincial improved seed base for *A. nepalensis* was established at the state-owned Yipinglang Forest Farm in Lufeng County, and Yubaiding Forest Farm in Eshan County, Yunnan Province, China. With this, the conservation of germplasm resources for *A. nepalensis* was now under development. However, all of the superior *A. nepalensis* trees in the germplasm bank were from Yunnan Province, with some of them even from the same town. Thus, there are likely close genetic relationships between various superior

© *Forest Research Institute Malaysia*

trees; specifically, there may be a high genetic redundancy of these resources in the germplasm bank. To reduce genetic redundancy, it is extremely urgent to create a high-quality core collection for the germplasm bank.

The key to the development of core collection for plants is to devise an optimal strategy, which primarily includes the optimal data type, sampling strategy, and evaluation method (Guruprasad et al. 2014, Zeng et al. 2014). For this study, a strategy for the development of a core collection for *A. nepalensis* (the original germplasm resources of which were from superior *A. nepalensis* trees in the germplasm bank) was discussed following that of a core collection for *P. yunnanensis* (the original germplasm resources of which were from natural forests in the overall *P. yunnanensis* distribution area) (Wang et al. 2019a, b, 2021). This research may provide an approach towards the creation of core collections for superior trees and clones obtained through breeding, in order to improve the conservation efficiencies of *A. nepalensis* germplasm resources.

MATERIALS AND METHODS

Original germplasm materials

A total of 84 samples from 28 superior *A. nepalensis* trees were investigated, all of which were from Yunnan Province, China (Table 1). Three samples of 11-year-old trees were randomly selected from the half-sibling progenies of each superior tree, which were planted in the provincial *A. nepalensis* germplasm bank being established at the state-owned Yubaiding Forest Farm in Eshan County. An amount of 10 g mature leaves of each sample were used for the extraction of genomic DNA and analysis of SRAP (sequence-related amplified polymorphism) markers. We immediately sorted the leaves of each sample as they were collected from the trees into paper bags containing silica gel for rapid dehydration at room temperature. Simultaneously, 30 ripe infructescences of each sample were collected to determine their phenotype traits using vernier callipers and ruler. Immediately after being removed from the trees, infructescences of each sample were sorted into yarn mesh bags. Once the infructescences were dehiscent by natural drying, the seeds were obtained, and the phenotype traits of the seeds were quantified using a pair of vernier callipers. The tree height (TH), diameter at breast height (DBH), east–west crown diameter (EWCD) (long crown diameter), and north–south crown diameter (NSCD) (short crown diameter) of each sample were determined separately.

Determination of phenotype characteristics and analysis of data

Indicators and techniques for the determination of the phenotype traits of trees were employed to measure 10 quantitative characteristics of the infructescences, seeds and vegetative growth (Boratynska et al. 2008, Xu 2015, Wang et al. 2021). For this study, the 10 quantitative characteristics included infructescence weight (IW), infructescence length (IL), infructescence diameter (ID), seed length (SL), seed width (SW), thousand seed weight (TSW), TH, DBH, EWCD and NSCD. The IL, ID, SL, and SW were determined using vernier callipers. An electronic balance was used to weigh IW and TSW. A total of 30 ripe infructescences from each sample and 30 seeds from each infructescence were randomly selected and their indicators were measured, after which the mean of each indicator was computed for each sample. DBH and TH of each sample were measured using diameter tape and laser height finder respectively. The decussation method was utilised to separately measure EWCD and NSCD of each sample using linen tape. The phenotype values were standardised on the basis of a standard deviation and divided into 10 grades (Xu et al. 2008, Liu et al. 2013, Wang et al. 2021). The intergroup linkage method and Euclidean genetic distance were employed for the phenotype similarity matrix of the samples (Wang et al. 2019a).

SRAP markers and data analysis

SRAP markers can detect the polymorphisms of open reading frames in a genome through a unique double primer design. Universal primers are employed as SRAP markers, thus, DNA samples can be directly amplified by PCR (polymerase chain reaction), and sample sequence data are not required. Hence, SRAP markers have been increasingly applied in the research of genetic diversity of plant germplasm resources, and the construction of a core collection (Budak et al. 2004, Wang et al. 2021). The reformed CTAB (cetyltrimethylammonium bromide) method (Porebski et al. 1997,

Table 1 Data on superior *Alnus nepalensis* trees and samples used in this research

Province of superior tree	Name of superior tree	Sampling number of half-sibling progeny for each superior tree	Longitude (E)	Latitude (N)
South of Yunnan	No. 2 of Menglie town of Jiangcheng County (JCML2)	3	101° 53'	22° 32'
	No. 2 of Mengxian town of Ninger County (NEMX2)	3	101° 15'	22° 56'
South-east of Yunnan	No. 2 of Wenliu town of Qiubei County (QBWL2)	3	104° 25'	24° 17'
	No. 1 of Jinping town of Qiubei County (QBJP1)	3	104° 14'	24° 4'
	No. 6 of Xiajinchang town of Malipo County (MLPXJC6)	3	104° 48'	23° 10'
	No. 2 of Jinping town of Qiubei County (QBJP2)	3	104° 14'	24° 4'
	No. 5 of Xiajinchang town of Malipo County (MLPXJC5)	3	104° 48'	23° 10'
	No. 2 of Zhela town of Yanshan County (YSZL2)	3	104° 28'	23° 40'
	No. 4 of Xiajinchang town of Malipo County (MLPXJC4)	3	104° 48'	23° 10'
	No. 5 of Zhujie town of Guangnan County (GNZJ5)	3	104° 56'	23° 44'
	No. 4 of Zhujie town of Guangnan County (GNZJ4)	3	104° 56'	23° 44'
	No. 1 of Zhela town of Yanshan County (YSZL1)	3	104° 28'	23° 40'
	No. 3 of Zhujie town of Guangnan County (GNZJ3)	3	104° 56'	23° 44'
	No. 2 of Xier town of Mile City (MLXE2)	3	103° 11'	24° 25'
	No. 3 of Zhela town of Yanshan County (YSZL3)	3	104° 28'	23° 40'
Central region of Yunnan	No. 5 of Xiaojie town of Yimen County (YMXJ5)	3	102° 8'	24° 51'
	No. 5 of Yangwu town of Xinping County (XPYW5)	3	102° 9'	23° 55'
	No. 3 of Xiaojie town of Yimen County (YMXJ3)	3	102° 8'	24° 51'
	No. 9 of Xiaojie town of Yimen County (YMXJ9)	3	102° 8'	24° 51'
	No. 3 of Pingdian town of Xinping County (XPPD3)	3	101° 57'	24° 3'
	No. 1 of Fuliangpeng town of Eshan County (ESFLP1)	3	102° 5'	24°18'
South-west of Yunnan	No. 1 of Gelanghe town of Menghai County (MHGLH1)	3	100° 34'	21°52'
	No. 1 of Yizhi town of Jinggu County (JGYZ1)	3	100° 35'	23°12'
	No. 6 of Yaoquyaozu town of Mengla County (MLYZ6)	3	101° 32'	21° 43'
	No. 3 of Mengsong town of Menghai County (MHMS3)	3	100° 34'	22° 3'
	No. 1 of Menghai town of Menghai County (MHMH1)	3	100° 26'	21° 59'
	No. 1 of Fuyan town of Menglian County (MLFY1)	3	99° 31'	22°28'
	No. 2 of Yongping town of Jinggu County (JGYP2)	3	100° 21'	23° 18'

Mi et al.2017) was used for genomic DNA extraction from the leaves of each sample. We read the primer sequences of SRAP marker according to Budak et al. (2004). In view of the number, clarity and repeatability of amplification bands, 28 pairs of primers were screened out for formal amplification (Table 2). We consulted Ablitif (2014) for the SRAP–PCR procedure and made some necessary changes, on which the reaction mixture system and amplification procedure for the research was established (Table 3). The amplification products of SRAP–PCR were tested using 8% polyacrylamide gel electrophoresis (Ablitif 2014). NTSYSpc2.10e and POPGENE32 software were employed to calculate the Nei's genetic distance and indices of genetic diversity respectively according to the SRAP marker data of the samples (Wang et al. 2019b).

Creation of core collection

The Euclidean genetic distance (Wang et al. 2019a) was computed on the basis of 10 phenotype trait values. Clustering was achieved using an unweighted average method with SPSS17.0 and NTSYSpc2.10e software. Six sampling fractions (i.e. 10, 15, 25, 35, 45 and 55%) were established to screen the optimal sampling fraction. For each cluster, when a sample had a minimum genetic distance from other samples, it was removed and not added to the core collection, which was the improved minimum distance stepwise sampling method (IMDSSM) adopted for this study (Xu 2005, Wang et al. 2021). Six germplasm subsets were established using the strategies above.

Nei's genetic distance (Wang et al. 2019b)

Table 2 Sequences of primers used in the SRAP–PCR system based on the results of screened primers

SRAP primer sequence				28 pairs of primers in the experiment
Number	Sequence	Number	Sequence	
Me1	TGAGTCCAAACCGGATA	Em1	GACTGCGTACGAATTAAT	Me1 + Em5, Me3 + Em7, Me1 + Em7 Me2 + Em2, Me6 + Em7, Me7 + Em5 Me9 + Em2, Me9 + Em4, Me9 + Em5 Me10 + Em4, Me10 + Em8, Me10 + Em9 Me1 + Em6, Me2 + Em3, Me3 + Em1 Me3 + Em10, Me4 + Em5, Me4 + Em6 Me4 + Em8, Me5 + Em1, Me5 + Em5 Me5 + Em9, Me6 + Em1, Me6 + Em5 Me6 + Em8, Me7 + Em8, Me9 + Em10 Me10 + Em10
Me2	TGAGTCCAAACCGGAGC	Em2	GACTGCGTACGAATTTGC	
Me3	TGAGTCCAAACCGGCAG	Em3	GACTGCGTACGAATTGAC	
Me4	TGAGTCCAAACCGGACC	Em4	GACTGCGTACGAATTTGA	
Me5	TGAGTCCAAACCGGAAG	Em5	GACTGCGTACGAATTAAC	
Me6	TGAGTCCAAACCGGTAA	Em6	GACTGCGTACGAATTGCA	
Me7	TGAGTCCAAACCGGTCC	Em7	GACTGCGTACGAATTCAA	
Me8	TGAGTCCAAACCGGTGC	Em8	GACTGCGTACGAATTCTT	
Me9	TGAGTCCAAACCGGAAC	Em9	GACTGCGTACGAATTGAG	
Me10	TGAGTCCAAACCGGTAG	Em10	GACTGCGTACGAATTGCC	

Table 3 Amplification reaction system and amplification program of the SRAP–PCR

Amplification reaction system		Amplification program
Composition	Dosage (μL)	
DNA (50 ng μL^{-1})	2	Initial denaturation at 94 °C for 5 min and 5 cycles of denaturation at 94 °C for 1 min, annealing at 35 °C for 1 min and extension at 72 °C for 1 min, followed by 35 cycles of denaturation at 94 °C for 1 min, primer annealing at 44 °C for 1 min, and extension at 72 °C for 1 min. The amplification process was completed with a 7 min final extension at 72 °C and the PCR products were maintained at 4°C
Forward primer(10 μmol L^{-1})	1.5	
Reverse primer(10 μmol L^{-1})	1.5	
MIX(MgCl$_2$ 2 mM, KCl 50 mM, 10 mM (NH$_4$)$_2$ SO$_4$, BSA 100 mg mL^{-1} T$_6$1U)	10	
ddH$_2$O	5	
Total	20	

was computed based on 201 polymorphic loci of SRAP markers data. The clustering was achieved via unweighted average method using POPGENE32 and NTSYSpc2.10e software. Six sampling fractions (i.e. 10, 15, 25, 35, 45 and 55%) were established to screen the optimal sampling fraction, and IMDSSM was adopted for sampling (Xu 2005, Wang et al. 2021). Six germplasm subsets were established using the strategies above.

The mixed genetic distance was computed on the basis of integrating 10 phenotype trait values and 201 polymorphic loci of SRAP markers data. The formulae for the mixed genetic distance, genetic distance of phenotype traits, and genetic distance of molecular markers are as reported by Liu et al. (2012) and Wang et al. (2021). Clustering was achieved via unweighted average method using Matlab (R2017a) software. Six sampling fractions were established to screen the optimal sampling fraction, and IMDSSM (Wang et al. 2021, Xu 2005) was adopted for sampling. Thus, six germplasm subsets were established using the strategies above.

Evaluation of core collection

Appraisal of the core collection was carried out to identify an optimum sample strategy and germplasm subset based on the representativeness of the core collection for the original germplasm collection through the indicators system. For this research, the quantitative characteristics consisted of 10 phenotype traits with continuous values. Meanwhile, the qualitative characteristics consisted of 201 polymorphic loci with discontinuous values. For the 18 germplasm subsets of this study, six were established using quantitative characteristics, six were established via qualitative characteristics, and the remaining six were established by combining quantitative and qualitative characteristics. In earlier research, the effective appraisal indices for the representativeness of a core collection for the original germplasm collection were classified into two categories (i.e. appraisal parameters of quantitative characteristics and appraisal parameters of qualitative characteristics). One of the categories was entirely appropriate for the appraisal of a core collection established by quantitative characteristics, while another was suitable through qualitative characteristics (Zhao & Zhang 2007, Zhang 2010). A coalition between the two categories was adopted to appraise a core collection established by combining quantitative and qualitative characteristics (Liu et al. 2012). Evaluation parameters of quantitative characteristics included the mean difference percentage, variance difference percentage, range coincidence rate, change rate of variation coefficient, retention ratio of phenotype trait, and Shannon-Weaver diversity indices (H') (Hu et al. 2000, Liu et al. 2013, Wang et al. 2019a). The evaluation parameters of qualitative characteristics included the number of alleles, retention rate of alleles, number of effective alleles, Nei's genetic diversity index (H), and Shannon's information index (I) (Wang et al. 2019b). The evaluation parameters of the quantitative characteristics and those of the qualitative characteristics were calculated using Excel 2012 and POPGENE 32 software respectively (Wang et al. 2021).

Confirmation of core collection

It is necessary to confirm core collection in order to verify the effectiveness of the optimum sample strategy and germplasm subsets through specific approaches. For this purpose, genetic distance comparison (Ablitif 2014, Wang et al.2019b) and clustering analysis (Wang et al. 2021) were employed to confirm the efficacy of the core collection being screened out using Matlab (R2017a) and POPGENE32 software separately.

RESULTS

Evaluation of quantitative traits of original germplasm collection and germplasm subsets

For each of the six developed germplasm subsets based on phenotype values and Euclidean genetic distance, the average values of H' for the 10-phenotype characteristics of a germplasm subset initially increased and then decreased with the diminution of the sampling fraction, of which the maximum was the sampling proportion at 45% (Table 4). The H' of the three germplasm subsets (2.030, 2.052, 2.039) at 55, 45 and 35% sampling proportions were greater than that of the original germplasm collection (1.986). However, those of the other three germplasm subsets (1.966, 1.826, 1.660) at 25, 15 and 10%

sampling proportions were lower than that of the original germplasm collection. For the six germplasm subsets being established through the integration of phenotype and molecular markers data and mixed genetic distance, the average H' values for the 10 phenotype characteristics of a germplasm subset decreased with the diminution of the sampling fraction (Table 4). The H' of the three germplasm subsets (2.027, 2.024, 1.990) at 55, 45 and 35% sampling proportions were greater than those of the original germplasm collection. However, those of the other three germplasm subsets (1.938, 1.762, 1.596) at 25, 15 and 10% sampling proportions were lower than that of the original germplasm collection. Consequently, the six germplasm subsets at 55, 45 and 35% sampling proportions from the two data types and two genetic distance methods were advantageous for promoting the genetic diversity indicators and decreasing genetic redundancy.

In this study, H' showed highly significant differences between the germplasm collections (i.e. the original germplasm collection and the 12 germplasm subsets being developed using six sampling proportions, two data types, and two genetic distance methods) through mean variance analysis. Furthermore, the average H' for the 10 phenotype characteristics of the original germplasm collection was significantly larger than that of the four germplasm subsets that were created based on two sampling proportions (15 and 10%), two data types and two genetic distance methods through mean multiple comparison analysis. However, no marked differences were observed between the original germplasm collection and the other eight germplasm subsets (Table 4). In summary, the eight germplasm subsets being constructed based on 55, 45, 35 and 25% sampling proportions, two data types and two genetic distance methods embodied the genetic diversity of the original germplasm collection.

Using the same sampling proportion, the average of H' for the 10 phenotype characteristics of the germplasm subsets created based on the phenotype values and Euclidean genetic distance was higher than that of the germplasm subsets developed through the integration of phenotype and molecular markers data and mixed genetic distance; however, no significant differences were observed between them (Table 4). To sum up, considering the H' of the germplasm subsets, time, economic costs, and the convenience of obtaining data for the development of the germplasm subsets, it was revealed that the strategy of phenotype values and Euclidean genetic distance was more appropriate than integrating the phenotype and molecular markers data and mixed genetic distance for the creation of a core collection for superior *A. nepalensis* trees.

For the 12 germplasm subsets (six were created based on phenotype values and Euclidean genetic distance, and the rest, developed by integrating phenotype and molecular markers data and mixed genetic distance), the mean difference percentages (0.39–9.77%) between the germplasm subsets and original germplasm collection were < 20%, and the coincidence rates (82.01–100%) between the germplasm subsets and original germplasm collection were > 80%. The retention ratios of phenotype traits (87.10–100%) of the eight germplasm subsets (i.e. 55, 45, 35 and 25% sampling proportions, two data types, and two genetic distance methods) were > 85%, while those (59.14–74.27%) of the other four germplasm subsets (i.e. 15 and 10% sampling fractions, two data types and two genetic distance methods) were < 75% (Table 5). For the four germplasm subsets created based on phenotype values and Euclidean genetic distance (i.e. 55, 45, 35 and 25% sampling proportions), the change rate of variation coefficient (125.84%) and variance difference percentage (21%) of the germplasm subset at a 25% sampling proportion were maximum. Similarly, for the four germplasm subsets developed by integrating phenotype and molecular markers data and mixed genetic distance (i.e. 55, 45, 35 and 25% sampling proportions), the change rate of variation coefficient (116.02%) and variance difference percentage (30%) of the germplasm subsets at a 25% sampling proportion were also maximum. Thus, the 25% sampling fraction was superior to the other five sampling fractions from the perspective of these five evaluation parameters (Table 5). With the evaluation results of quantitative characteristics, we found that the germplasm subset created based on phenotype values and Euclidean genetic distance at a 25% sampling fraction was optimal in terms of the representativeness, effectiveness, and practicability for the development of a core collection for superior *A. nepalensis* trees.

Table 4 Comparison of Shannon-Weaver diversity indices for phenotypic traits between the original germplasm collection and germplasm subsets of *Alnus nepalensis*

Trait	Original germplasm collection	Germplasm subsets being constructed via phenotypic values and Euclidean genetic distance						Germplasm subsets being constructed via integrating phenotypic values and SRAP marker data, and mixed genetic distance					
		GS55	GS45	GS35	GS25	GS15	GS10	GS55	GS45	GS35	GS25	GS15	GS10
DBH (cm)	2.002	1.997	2.039	1.948	1.823	1.626	0.900	2.004	1.921	1.924	1.691	1.519	1.213
TH (m)	1.971	1.898	1.923	1.939	1.951	1.738	1.494	1.856	1.819	1.733	1.716	1.479	1.494
LCD (m)	2.075	2.062	2.057	2.079	2.083	1.925	1.906	2.035	1.997	1.855	1.832	1.672	1.494
SCD (m)	2.026	1.944	1.952	2.013	2.095	1.845	1.733	1.996	1.958	1.909	2.004	1.845	1.560
SL (cm)	1.964	2.091	2.118	2.034	1.925	1.626	1.733	2.086	2.120	2.018	1.910	1.586	1.494
SW (cm)	2.011	2.155	2.187	2.139	2.070	2.032	1.733	2.126	2.151	2.106	2.055	2.098	1.733
WTS (g)	1.793	1.866	1.906	1.897	1.686	1.672	1.733	1.907	1.947	1.986	1.950	1.733	1.494
IW (g)	2.025	2.118	2.111	2.145	2.112	1.992	1.733	2.070	2.064	2.064	2.112	1.992	1.906
IL (cm)	2.015	2.103	2.125	2.062	1.992	1.925	1.906	2.091	2.112	2.134	2.042	1.885	1.667
ID (cm)	1.978	2.068	2.104	2.139	1.923	1.885	1.733	2.104	2.152	2.166	2.070	1.818	1.906
Average H'	1.986 ± 0.075A	2.030 ± 0.099A	2.052 ± 0.096A	2.039 ± 0.090A	1.966 ± 0.136AB	1.826 ± 0.151BC	1.660 ± 0.290DE	2.027 ± 0.088A	2.024 ± 0.113A	1.990 ± 0.136A	1.938 ± 0.149AB	1.762 ± 0.203CD	1.596 ± 0.213E

GS55, GS45, GS35, GS25, GS15 and GS10 = germplasm subset at 55, 45, 35, 25, 15 and 10% sampling proportion respectively, DBH = diameter at breast height, TH = tree height, LCD = long crown diameter, SCD = short crown diameter, SL = seed length, SW = seed width, WTS = thousand seeds weight, IW = infructescence weight, IL = infructescence length, ID = infructescence diameter; capitals of A and B labels significantly different at 0.01 probability level, mean variance analysis indicated p = 0.000

Table 5 Comparison of the effective evaluation parameters of germplasm subsets of *Alnus nepalensis* based on phenotypic traits

Evaluation parameter	Germplasm subsets being constructed via phenotypic values and Euclidean genetic distance						Germplasm subsets being constructed via integrating phenotypic values and SRAP marker data, and mixed genetic distance					
	GS55	GS45	GS35	GS25	GS15	GS10	GS55	GS45	GS35	GS25	GS15	GS10
Phenotypic retention ratio (%)	100	100	97.85	89.25	74.27	62.37	98.92	98.92	92.47	87.10	72.04	59.14
Mean difference percentage (%)	9.62	9.77	9.63	8.48	6.34	5.72	0.96	0.86	0.53	0.39	3.80	6.97
Rate of variable coefficient change (%)	111.76	116.78	122.58	125.84	134.27	139.11	110.63	112.89	113.31	116.02	127.99	135.47
Coincidence rate of range (%)	100	100	98.95	97.16	94.23	85.94	99.91	98.72	94.53	90.15	87.58	82.01
Variance difference percentage (%)	0.00	0.00	20.00	21.00	20.00	20.00	0.00	0.00	10.00	30.00	40.00	40.00

GS55, GS45, GS35, GS25, GS15 and GS10 = germplasm subset at 55, 45, 35, 25, 15 and 10% sampling proportion respectively

Evaluation of qualitative traits of original germplasm collection and germplasm subsets

The number of polymorphic loci, percentage of polymorphic loci, number of alleles, and retention rate of alleles of the germplasm subsets, developed using molecular markers data and Nei's genetic distance, and by integrating phenotype and molecular markers data and mixed genetic distance, decreased with the diminution of the sampling fraction. Further, with the same sample proportion, the above indicators of the six germplasm subsets being developed based on molecular markers data and Nei's genetic distance were greater than, or equal to, that of the six germplasm subsets created by integrating the phenotype and molecular marker data and mixed genetic distance. The number of alleles of the original germplasm collection was significantly larger than that of the three germplasm subsets (one germplasm subset created based on molecular markers data and Nei's genetic distance at a 10% sampling proportion, and the other two were developed by integrating the phenotype and molecular marker data and mixed genetic distance at sampling proportions of 15 and 10%) through variance and multiple comparison analyses, however, no marked differences were observed between the original germplasm collection and the other nine germplasm subsets (Table 6).

The number of effective alleles, Nei's genetic diversity index, Shannon's information index, and population genetic diversity of the six developed germplasm subsets based on molecular markers data and Nei's genetic distance were greater than that of the original germplasm collection. These indicators initially increased and then decreased with the reduction of the sampling fraction. Further, through variance and multiple comparison analyses, the genetic diversity indices of the four germplasm subsets created based on molecular markers data and Nei's genetic distance at sampling proportions of 45, 35, 25 and 15% were significantly larger than those of the original germplasm collection (Table 6). The genetic diversity indices of all six germplasm subsets developed by integrating phenotype and molecular markers data and mixed genetic distance also initially increased and then decreased with the decreased sampling fraction. All of the above germplasm subset indices at a 10% sampling proportion were lower than those of the original germplasm collection. Furthermore, the Shannon's information index of the germplasm subset created by integrating phenotype and molecular markers data and mixed genetic distance at 10% sampling fraction was significantly lower than that of the original germplasm collection as shown by the variance and multiple comparison analyses (Table 6). With the same sampling proportion, these genetic diversity indices of germplasm subsets developed through the integration of phenotype and molecular markers data and mixed genetic distance were lower than that of the germplasm subsets created based on molecular markers data and Nei's genetic distance. With the evaluation results of qualitative characteristics, it was found that the germplasm subset created using molecular markers data and Nei's genetic distance at 15% sampling fraction was optimal in terms of the representativeness, effectiveness, and practicability for the development of a core collection for superior *A. nepalensis* trees.

In conclusion, for the 18 germplasm subsets (six germplasm subsets created based on phenotype values and Euclidean genetic distance, six germplasm subsets developed using molecular markers data and Nei's genetic distance, and the remaining six germplasm subsets generated using combined phenotype and molecular markers data and mixed genetic distance), the sampling proportions of 55, 45, 35, and 25% could represent the original germplasm collection, as well as those based on molecular markers data and Nei's genetic distance at a 15% sampling proportion. In particular, molecular markers data, Nei's genetic distance, and a 15% sampling proportion were the elements of an optimal strategy for the establishment of a core collection for superior *A. nepalensis* trees.

Verification of core collection

All minimum and average genetic distances of the germplasm subset, created using molecular markers data and Nei's genetic distance at a 15% sampling fraction, were higher than those of the original germplasm collection. The germplasm subset increased these indicators by 235.97 and 40.04% respectively (Table 7) Thus, several redundant original germplasm collection samples could be removed. The germplasm subset and original germplasm collection clustering revealed that the genetic distances between the germplasm

Table 6 Results of genetic diversity between original germplasm collection and germplasm subsets of *Alnus nepalensis*

Evaluation parameter	Original germplasm collection	Germplasm subsets constructed via SRAP markers data and Nei's genetic distance (D_m)						Integrating the phenotypic traits and SRAP markers data and mixed genetic distance (D_{mix})					
		GS55	GS45	GS35	GS25	GS15	GS10	GS55	GS45	GS35	GS25	GS15	GS10
QG	84	46	38	29	21	13	8	46	38	29	21	13	8
NPL	201	201	201	201	201	200	191	201	200	200	198	194	188
PPL%	100	100	100	100	100	98.51	95.02	100	99.50	99.50	98.51	96.52	93.53
Na	2.00	2.00	2.00	2.00	2.00	1.98	1.95*	2.00	2.00	2.00	1.99	1.97*	1.94*
RRA%	100	100	100	100	100	99.20	97.50	100	100	100	99.50	98.50	96.80
Ne	1.662	1.704	1.718*	1.722*	1.730*	1.731*	1.701	1.692	1.696	1.698	1.691	1.686	1.645
H	0.379	0.397	0.402*	0.404*	0.409*	0.405*	0.393	0.393	0.399	0.394	0.390	0.386	0.367
I	0.559	0.577*	0.586*	0.589*	0.594*	0.588*	0.569	0.575	0.580	0.576	0.570	0.563	0.539*
Ht	0.379	0.397	0.402*	0.404*	0.409*	0.405*	0.393	0.393	0.399	0.394	0.390	0.386	0.367

QG = quantity of germplasm, NPL = number of polymorphic loci, PPL = percentage of polymorphic loci, Na = number of alleles, RRA = retention rate of alleles, Ne = number of effective alleles, H = Nei's genetic diversity index, I = Shannon's information index, Ht = populations genetic diversity, * = significantly different at 0.05 probability level

Table 7 Comparison of genetic distances between the original germplasm collection and the core collection (molecular markers data and Nei's genetic distance at a 15% sample proportion) of *Alnus nepalensis*

Population of germplasm	Quantity of germplasm	Min genetic distance	Max genetic distance	Average genetic distance
Original germplasm collection	84	0.139	0.826	0.487
Core collection (15% sampling proportion)	13	0.467	0.826	0.682

subset samples were higher than those of the original germplasm collection (Figures 1 and 2). The results revealed that the germplasm subset, developed using molecular markers data and Nei's genetic distance at a 15% sampling proportion, was an efficient core collection for superior *A. nepalensis* trees.

DISCUSSION

Data types for establishing a core collection for superior *A. nepalensis* trees

Three types of data were employed for the establishment of a core collection, which included agronomic morphological traits, molecular markers and combined phenotype and molecular markers data. Liu et al. (2012) compared three data types, namely, agronomic morphological characteristics data (15 agronomic traits of leaves, flowers, and fruits) and Euclidean genetic distance, SSR molecular markers data and Nei's genetic distance, and integrated phenotype and SSR molecular markers data and mixed genetic distance. They observed that the representativeness of the core collection developed by integrating phenotype and SSR molecular markers data and mixed genetic distance was superior to that of the core collection created based on agronomic morphological characteristics data and Euclidean genetic distance, or the SSR molecular markers data and Nei's genetic distance alone.

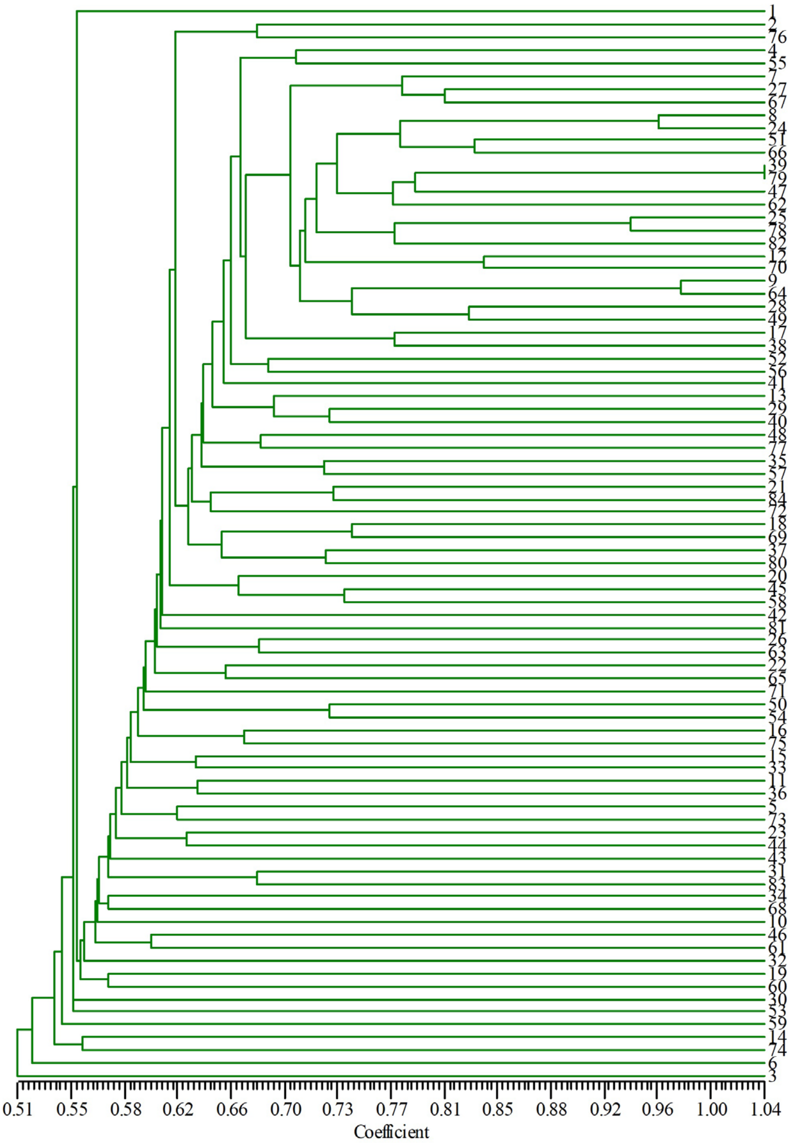

Figure 1 Cluster analysis map of the original germplasm collection (molecular markers data and Nei's genetic distance) for *Alnus nepalensis*

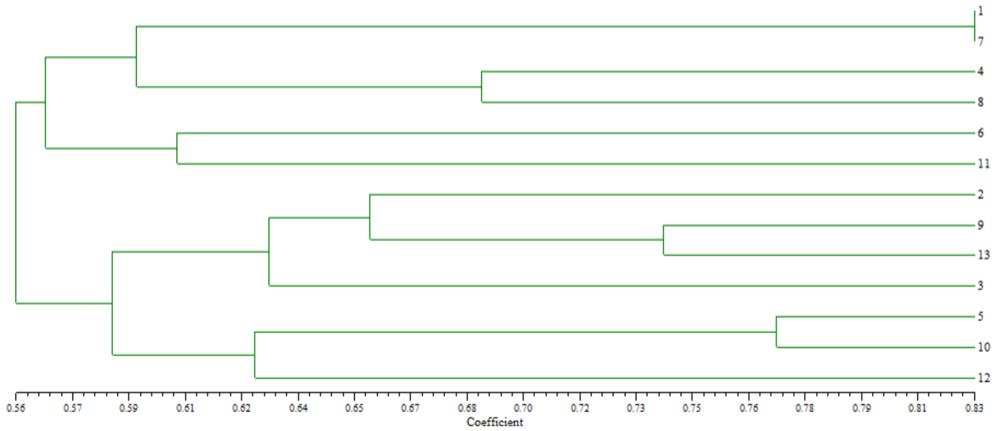

Figure 2 Cluster analysis map of the core collection (molecular markers data and Nei's genetic distance at 15% sample proportion) for *Alnus nepalensis*

We used 780 samples from natural *P. yunnanensis* populations as the original germplasm collection to explore the optimal development strategy of its core collection. The three data types, including agronomic morphological characteristics data (18 agronomic traits of needle, fascicle, cone, seed, stem and crown) and Euclidean genetic distance, SRAP molecular markers data (669 polymorphic loci) and Nei's genetic distance, and the integrated phenotype and SRAP molecular markers data and mixed genetic distance, were compared by evaluating the representative of the core collection for the original germplasm collection. The results indicated that the representativeness of the core collection created by integrating the phenotype and SRAP molecular markers data and mixed genetic distance was better than that of the core collection developed using agronomic morphological characteristics data and Euclidean genetic distance, or SRAP molecular markers data and Nei's genetic distance alone (Wang et al. 2021). Results show that the representativeness of the core collection created using SRAP molecular markers data and Nei's genetic distance was better than that of the core collection developed using agronomic morphological characteristics data and Euclidean genetic distance, and combined phenotype and SRAP molecular markers data and mixed genetic distance. In summary, the optimal data types being utilised to create the core collection for different plant germplasm resources are not always the same. They change with plant species, sample size, and genetic diversity of the original germplasm collection, as well as with the genetic distance calculation method. Thus, the optimal data type needs to be screened according to the specific study.

Sampling fraction for establishment of a core collection for superior *A. nepalensis* trees

An optimal sampling proportion was critical for the core collection development. There was more genetic redundancy in a core collection when the sampling proportion was excessively high. Therefore, the practicability and effectiveness of the core collection were reduced. On the contrary, some important germplasm materials could be lost when the sampling proportion was too low. In general, a relatively high sampling proportion is required if the original germplasm collection has a small sample size and high genetic diversity, while a relatively low sampling proportion is required if the original germplasm collection has large sample size and low genetic diversity. Thus, the sampling proportion is 30%, which is typically ~10% (Li et al. 2002). To give a few examples, a total of 161 samples from *Manihot esculenta* were used as original germplasm collection, and the optimal sampling proportion for the creation of its core collection was 15% (Wei et al. 2016); 45 samples from *Capsicum annuum* were used as the original germplasm collection, and the optimal sampling fraction for the development of its core collection was 30% (Jiao et al. 2018); and 1197 samples from *Citrullus lanatus* were used as original germplasm collection, and the optimal sampling fraction of its core collection creation was 10.9% (Zhang et al. 2016). Other such examples can be found in Yuan (2012), Zhang et al. (2019) and Wang et al. (2021).

For this study, a total of 84 samples from *A. nepalensis* were used as the original germplasm collection, and the optimal sampling fraction for its core collection creation was 15%. This fell within the generally accepted range of sampling fraction for the development of a core collection for plants. Of course, the optimal sampling proportion depended on the sample size and genetic diversity of the original germplasm collection, and the sampling strategy (i.e. data types, methods for calculating genetic distance, sampling methods, and clustering methods) used in the core collection development.

CONCLUSION

In this study, 18 germplasm subsets were created based on three data types (i.e. phenotype values, SRAP molecular markers data, and combined phenotype and SRAP molecular markers data), three methods for calculating genetic distance (i.e. Euclidean genetic distance, Nei's genetic distance, and mixed genetic distance), six sampling proportions (i.e. 55, 45, 35, 25, 15, and 10%), IMDSSM, and unweighted average method. The effectiveness and representativeness of the 18 germplasm subsets for the original germplasm collection were evaluated by the indicators of quantitative and qualitative characteristics. The results reveal that 13 germplasm subsets can represent the original germplasm collection, of which four germplasm subsets were developed based on phenotype values and Euclidean genetic distances at sampling proportions of 55, 45, 35 and 25%. A total of five germplasm subsets were created based on SRAP molecular markers data and Nei's genetic distance at sampling proportions of 55, 45, 35, 25 and 15%, while four germplasm subsets were developed using integrated phenotype and SRAP molecular markers data and mixed genetic distance at sampling proportions of 55, 45, 35 and 25%. Using the same sampling proportions, the germplasm subsets created using SRAP molecular markers data are more representative of the original germplasm collection than those created based on two other data types. In conclusion, we obtained an optimal sampling strategy for the development of a core collection for superior *A. nepalensis* trees, which primarily included SRAP molecular markers data, Nei's genetic distance, a 15% sampling proportion, unweighted average method and IMDSSM. This sampling strategy may be employed to establish core collections for superior trees and clones obtained through breeding projects.

ACKNOWLEDGEMENT

We acknowledge Wang LG for Matlab software programming to calculate the mixed genetic distances, Zhang WF for creating the figures of the sample location data, and the professors of the Southwest Forestry University in China. This work was supported by the grants from open fund project of Key Laboratory for Forest Resources Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University (KLESWFU-201905). We would also like to thank Cao SM for careful reading of the manuscript.

REFERENCES

Ablitif Y. 2014. Evaluation on germplasm resources and structure of core germplasm bank of Sinkiang pear (*Pyrus* L.). PhD thesis, Xin Jiang Agricultural University, Urumqi.

Arriel DAA, Fajardo CG, Vieira FDA & De CD. 2023. Spatial genetic structure of sympatric populations of *Eremanthus* species in Brazil: implications for management. *Journal of Tropical Forest Science* 35: 56–65. https://doi.org/10.26525/jtfs2023.35.1.56

Boratynska K, Jasiriska AK & Ciepluch E. 2008. Effect of tree age on needle morphology and anatomy of *Pines uliginosa* and *Pines silvestris*-species-specific character separation during ontogenesis. *Flora* 203: 617–626. https://doi.org/10. 1016/j.flora.2007.10.004

Budak H, Shearman RC, Gaussin RE & Dweikat I. 2004. Application of sequence-related amplified polymorphism markers for characterization of turf grass species. *HortScience* 39: 955–958. https://doi.org/10.21273/HORTSCI.39.5.955

Frankel OH. 1984. Genetic perspectives of germplasm conservation. Pp 161–170 in Arber WK et al. (eds) *Genetic Manipulation: Impact on Man and Society*. Cambridge University Press, Cambridge.

Guardo DM, Scollo F, Ninot A et al. 2019. Genetic structure analysis and selection of a core collection for carob tree germplasm conservation and management. *Tree Genetics & Genomes* 15: 41–47. https://doi.org/10.1007/s11295-019-1345-6

Guruprasad R, Krishnan RR, Dandin SB & Girish NV. 2014. Groupwise sampling: a strategy to sample core entries from RAPD marker data with application to mulberry. *Trees* 28: 723–731. https://doi.org/10.1007/s00468-014-0984-3

Hu J, Xu HM & Zhu J. 2000. Constructing core collection of crop germplasm by multiple clusters based on genotypic values. *Journal of Biomathematics* 15: 103–109.

Jiao YS, Ren FS, Guo ZW, Chen HF, Liu HJ & Sun Q. 2018. Genetic diversity analysis and special core collection construction of spiral pepper germplasms. *Journal of Henan Agriculture Science* 47: 99–105. https://doi.org/10.15933/j.cnki.1004-3268.2018.09.017

Li DB, Xu N, Qin XQ et al. 2020. Genetic diversity and construction of core collections of litchi (*Litchi chinensis* Sonn.) germplasm originated and introduced in Guangxi. *Journal of Southern Agriculture* 51: 1537–1544.

Li ZC, Zhang HL, Zeng YW et al. 2002. Studies on sampling schemes for the establishment of core collection of rice landraces in Yunnan, China. *Genetic Resources and Crop Evolution* 49: 67–74. https://doi.org/10.1023/A:1013855216410

Liu DH, Zhang FQ & Zhang WH. 2013. Establishment of *Eucalyptus urophylla* core collection based on geographical distribution and phenotypic data. *Journal of Southwest Forestry University* 33: 1–8.

Liu ZC, Liu DL, Cui M et al. 2012. Combining agronomic traits and molecular marker data for constructing *Malus sieversii* core collection. *Acta Horticulturae Sinica* 39: 1045–1054. https://doi.org/10.16420/j.issn.0513-353x.2012.06.006

Mi FT, Liu J & Li YY. 2017. Studies on extraction method of genomic DNA from leaves of *Alnus nepalensis* D.Don. *Tropical Agriculture Science Technology* 40: 35–37.

Porebski S, Bailey LG & Baum BR. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* 15: 8–15. https://doi.org/10.1007/BF02772108

Wang XL, Gao CJ & Li K. 2019a. Construction strategy and verification of timber-used germplasm conservation bank of *Pinus yunnanensis. Journal of Plant Resources and Environment* 28: 105–114. https://doi.org/10.3969/j.issn.1674-7895.2019.01.13

Wang XL, Gao CJ & Li K. 2019b. Strategy for constructing *Pinus yunnanensis* germplasm bank for timber based on the SRAP molecular marker. *Plant Science Journal* 37: 211–220.

Wang XL, Cao ZL, Gao CJ & Li K. 2021. 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* 68: 3219–3240. https:// doi.org/10.1007/s10722-021-01182-9

Wei ZS, Fu HT & Tian YN. 2016. Construction of the core collection of *Manihot esculenta* Crantz. *Journal of Anhui Agriculture Science* 44: 22–25. https://doi.org/10.13989/j.cnki.0517-6611.2016.10.009

Xu HM. 2005. Study on methods of constructing core collection of germplasm and their applications in core construction. PhD thesis, Zhe Jiang University, Hangzhou.

Xu N, Cheng XZ, Wang SH, Wang LX & Zhao D. 2008. Establishment of an Adzuki bean (*Vigna angularis*) core collection based on geographical distribution and phenotypic data in China. *Acta Agronomica Sinica* 34: 1366–1373. https://doi. org/10.3724/SP.J.1006.2008.01366

Xu YL. 2015. Genetic Variation of Natural Populations in *Pinus yunnanensis* Franch. PhD thesis, Bei Jing Forestry University, Haidian

Yuan HT. 2012. Construction of Xinjiang wild walnut germplasm resource basic database and research on methods of building core collection. PhD thesis, Xin Jiang Agricultural University, Urumqi.

Zeng XJ, Li D, Hu YP, Huang QJ & Su XH. 2014. A preliminary study on construction of high-quality core collection of *Populus nigra. Scientia Silvae Sinicae* 50: 51–58.

Zhang D. 2010. *Genetic* diversity and core collection of wild castor-oil plant in South China. PhD thesis, Guangdong Ocean University, Zhanjiang.

Zhang HM, Zhai W, Jin HJ, Ding XT & Yu JZ. 2019. Genetic diversity analysis of 23 cucumber germplasms and screening of core germplasm resources using InDel markers. *Acta Agriculturae Shanghai* 35: 28–33. https://doi.org/10.15955/j.issn1000-3924.2019.04.05

Zhang HY, Fan JG, Guo SG, Ren Y & Gong GY. 2016. Genetic diversity, population structure, and formation of a core collection of 1197 *Citrullus* accessions. *HortScience* 51: 23–29. https://doi.org/10.21273/HORTSCI.51.1.23

Zhang J, Zhang P & Li Qx. 2018. Construction of core germplasm of Xinjiang wild walnut. *Journal of Fruit Science* 35: 168–176. https://doi.org/10.13925/j.cnki.gsxb.20170342

Zhao B & Zhang QX. 2007. Preliminary construction of the core germplasm of *Chimonanthus praecox* in China. *Journal of Beijing Forestry University* SI: 16–21.