SCIENTIFIC REPORTS
natureresearch

OPEN

Genetic diversity, phylogenetic structure and development of core collections in *Melilotus* accessions from a Chinese gene bank

Received: 11 March 2019
Accepted: 14 August 2019
Published online: 10 September 2019

Hongxiang Zhang[1,2], Rong Bai[1], Fan Wu[1], Wenli Guo[1], Zhuanzhuan Yan[1], Qi Yan[1], Yufei Zhang[1], Jinxing Ma[3] & Jiyu Zhang[1]

Melilotus **is an important forage legume, with high values as feed and medicine, and widely used as green manure, honey plant, and wildlife habitat enhancer. The genetic diversity, structure and subdivision of this forage crop remain unclear, and plant genetic resources are the basis of biodiversity and ecosystem diversity and have attracted increasing attention. In this study, the whole collection of 573 accessions from the National Gene Bank of Forage Germplasm (NGBFG, China) and 48 accessions from the National Plant Germplasm System (NPGS, USA) in genus *Melilotus* were measured with respect to five seed characters: seed length, width, width-to-length ratio, circumference and 100-seed weight. Shannon' genetic diversity index (H') and phenotypic differentiation (Pst) were calculated to better describe the genetic diversity. The ITS and *mat*K sequences were used to construct phylogenetic trees and study the genetic relationships within genus *Melilotu*. Based on seed morphology and molecular marker data, we preliminarily developed core collections and the sampling rates of *M. albus* and *M. officinalis* were determined to be 15% and 25%, respectively. The results obtained here provide preliminary sorting and supplemental information for the *Melilotus* collections in NGBFG, China, and establish a reference for further genetic breeding and other related projects.**

Melilotus is a forage legume of family, including 19 annual and biennial species, and three of the species have been cultivated: *M. albus*, *M. officinalis*, and *M. indicus*[1–3]. In comparison with most other forages, *Melilotus* has the advantages of tolerating extreme environmental conditions and providing high seed yields[4,5]. The nitrogen fixation rate of *Melilotus* is superior to those of other legumes, and it is beneficial in crop rotations[6]. Additionally, *Melilotus* can be used as a crop fertilizer[7] as well as nectar plants[8] and has important medicinal value due to the biological activity of their coumarins, which have many biological and pharmacological activities, including anti-HIV and anti-tumor effects[9]. During the past few years, *Melilotus*, as a good leguminous forage, has received much attention[10,11]. Plant genetic resources are the most essential of the world's natural resources and are of paramount importance for genetic improvement, germplasm innovation, and plant biology research; they play an important role in guaranteeing the food and nutrition security of an increasing population[12,13]. Abundant genetic resources have great potential to provide novel beneficial genes[14].

During the last 3–4 decades, major advances have been made in conserving these resources[15,16]. Although a large number of plant germplasm materials have been conserved in gene banks, their use is limited because of their overwhelming amount and lack of management[17]. According to Food and Agriculture Organization (FAO) estimates, only 1 million to 2 million of the 7.4 million germplasm accessions are specific and non-repetitive, while the remaining germplasm accessions contain different levels of repetition. An assessment and classification of the diversity is essential for effective utilization of the germplasm, and core germplasm development has

[1]State Key Laboratory of Grassland Agro-ecosystems; Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs; College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, 730020, P.R. China. [2]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, P.R. China. [3]National Quality Control & Inspection Centre for Grassland Industry Products, National Animal Husbandry Service, Ministry of Agriculture, Beijing, P.R. China. Hongxiang Zhang, Rong Bai and Fan Wu contributed equally. Correspondence and requests for materials should be addressed to J.Z. (email: zhangjy@lzu.edu.cn)

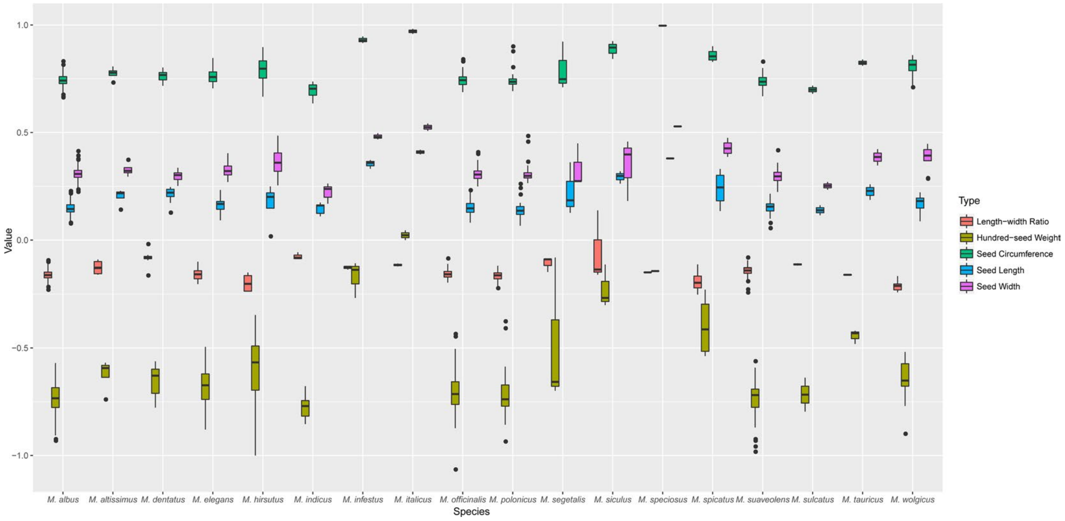

Figure 1. Morphologic variation analysis of five seed traits for 18 species. We calculated the logarithm of the values of five seed traits as ordinate in the box plot. Different traits are shown in different colors.

been proposed for better management and use of the collections available in gene banks[18,19]. A core collection can be defined as a minimum set of accessions representing maximum genetic diversity, and collections of the core set are described accurately and evaluated and managed carefully, for better conservation and utilization of germplasm accessions[20]. The common method of constructing a core set is to group the whole collection by morphological or molecular characteristics, then selecting the representative core accessions to form subcore groups and combining all subcore groups to construct the final core set[21,22]. The described core accessions could be more efficiently used for pre-breeding, genomic studies and conservation programs in gene banks.

Here, a total of 621 accessions of 18 *Melilotus* species, including the whole collection of 573 accessions from NGBFG, China, and 48 accessions from NPGS, USA, was analyzed to present a comprehensive view of the genetic diversity and phylogenetic structure among these accessions and provide the basis for constructing a core germplasm set. In our previous study, we selected 199 accessions to assess the genetic diversity in *Melilotus* and gain an initial understanding[23]. Seed morphology and the sequences of ITS and *mat*K were adopted to analyze genetic diversity and form core collections of *Melilotus*. Using seed traits to assess genetic diversity in the germplasm is advantageous in comparison with the use of other plant organs, as seeds are easy to collect and store[24]. More importantly, seed morphological traits can be utilized for species identification as well as selection criteria in crop improvement programs[25,26]. The nuclear DNA ITS and chloroplast DNA *mat*K have been widely applied in studies of inferring phylogenetic relationships at lower taxonomic levels and have been successfully used to analyze plant systematics[27–29]. The previous studies in Fabaceae indicated that the rate and pattern of ITS sequence mutation are appropriate for resolving relationships among species and genera[30], as well as revealed that *mat*K sequence can be used in phylogenetic analyses to successfully resolve relationships even at the species level[31]. Additionally, these sequences showed high stability and discrimination in *Melilotus*[32]. Examining both sequences and seed morphology might be an efficient method to analyze variation among *Melilotus* accessions and construct core sets.

Results

Seed morphological characterization.

The morphologic traits in seeds are presented in Fig. 1 and Supplementary Table S1. The mean values of seed length, width, width-to-length ratio, circumference and 100-seed weight were 2.332 cm, 1.694 cm, 0.723, 6.564 cm and 0.365 g, respectively. According to Supplementary Table S1, an analysis of variance indicated significant ($p < 0.05$) differences among species, but the values of all traits overlapped a lot in range for many species (Fig. 1). The box plot revealed the relationships of seed size and shape of 18 species as well as indicated a small number of outliers. What's more, we calculated the Pst parameter to assess the traits variation among species and the width-to-length ratio showed the lowest variation, while the 100-seed weight revealed the largest variation (the CV was 0.676 and the Pst parameter was 0.8473). The 100-seed weight and seed circumferences of *M. italicus*, *M. infestus*, *M. siculus* and *M. speciosus* were larger than those of the other species. Comparing the values of width-to-length ratio, circumference and 100-seed weight, the change tendencies of the latter two traits were similar since both two measures showed a positive correlation and reflected seed size. The width-to-length ratio was linked with the shape, and the difference among species was relatively small. Moreover, the CV values among species were larger than those within species, except for certain traits in a few species (the width-to-length ratios of *M. hirsutus* and *M. spicatus*, the circumference and 100-seed weight of *M. segetalis*).

Cluster analysis.

A total of 1145 sequences were newly amplified for this study. The nuclear DNA ITS sequences were successfully amplified for all 621 accessions, and the *mat*K sequences also performed well, with a

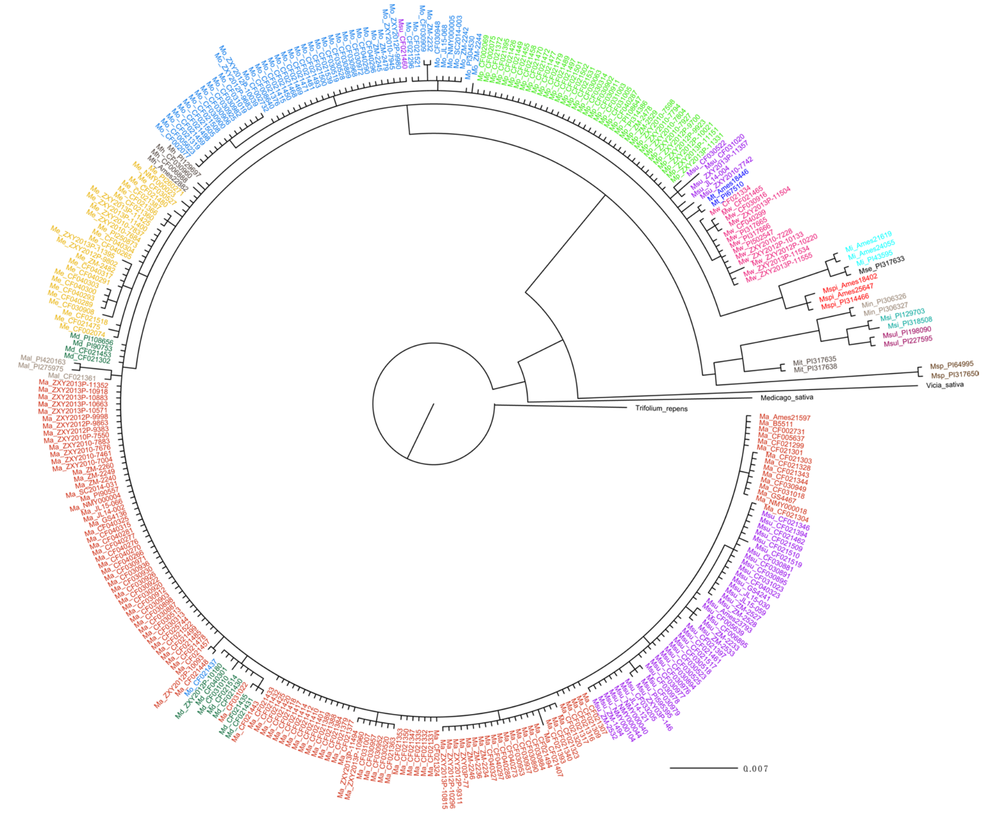

Figure 2. Bayesian tree of 18 species in *Melilotus* with branch lengths, based on ITS sequences. The abbreviations represent 18 species: Ma—*M. albus*, Mal—*M. altissimus*, Md—*M. dentatus*, Me—*M. elegans*, Mh—*M. hirsutus*, Mi—*M. indicus*, Min—*M. infestus*, Mit—*M. italicus*, Mo—*M. officinalis*, Mp—*M. polonicus*, Mse—*M. segetalis*, Msi—*M. siculus*, Ms—*M. speciosus*, Mpi—*M. spicatus*, Msu—*M. suaveolens*, Msul—*M. sulcatus*, Mt—*M. tauricus*, and Mw—*M. wolgicus*. See Supplement Table S3 for accession numbers.

high amplification rate of 99.3%. Based on these sequences, we constructed four phylogenetic trees to analyze the genetic diversity and phylogenetic structure of 18 species in *Melilotus*.

A phylogenetic tree of 18 species based on ITS sequences is shown in Fig. 2, with *Vicia sativa*, *Medicago sativa* and *Trifolium repens* as outgroups. Most species showed distinct diversity, and the result was similar to that of the previous study, in which 18 species formed two groups[23]. Ten species, which were *M. albus*, *M. suaveolens*, *M. altissimus*, *M. dentatus*, *M. elegans*, *M. hirsutus*, *M. officinalis*, *M. polonicus*, *M. tauricus* and *M. wolgicus*, formed a clade as the first group, and the others formed the second group. Most species showed small intraspecific distances, and several species, including *M. albus*, *M. suaveolens* and *M. dentatus*, have a very close genetic relationship. Nevertheless, not all accessions of *M. suaveolens* gather in a subclade, since several accessions came together with *M. polonicus*. It might be caused by gene flow and the pervious study indicated *M. suaveolens* could successfully crossed with *M. albus* and *M. polonicus*[33]. In contrast, the *mat*K sequences didn't perform well in assessing phylogenetic relationships in interspecific level. The diversity among 18 species revealed by *mat*K sequences was smaller, especially in the species *M. albus*, *M. altissimus*, *M. elegans*, *M. officinalis*, *M. polonicus*, *M. suaveolens* and *M. wolgicus*, as shown by their similar branch lengths (Supplementary Fig. S1), expect several accessions revealed variation with other accessions of the same species. The genetic diversity and relations could be reflected by the phylogenetic trees visually.

Additionally, *M. albus* and *M. officinalis* have been widely cultivated, and both species have been studied many times[34]. We selected these two species to construct phylogenetic trees (Fig. 3 and Supplementary Fig. S2) to assess the genetic diversity exactly and create a reference for developing core collections. Nearly all accessions are divided by species, which provided additional evidence about *M. albus* and *M. officinalis* should be treated as genetically isolated taxa. Both two species have low intraspecific genetic diversity, and the trees that contained only these two species could reflect their diversity more effectively. Based on the ITS and *mat*K trees, the same species grouped together except several individual materials. The results showed small genetic distance within the species, and most accessions had the same branch lengths.

Development of core collections. Two species, *M. albus* and *M. officinalis*, which stored large numbers of accessions in NGBFG, were selected to develop a representative core set. To determine an appropriate sampling ratio, six sampling proportions, 5%, 10%, 15%, 20%, 25% and 30%, were studied in our study. It is suggested that the coincidence rate (CR%) of range and the variable rate (VR%) for the coefficient of variation could evaluate the property of core collections[35]. We tried two different sampling methods, multiple clustering random sampling

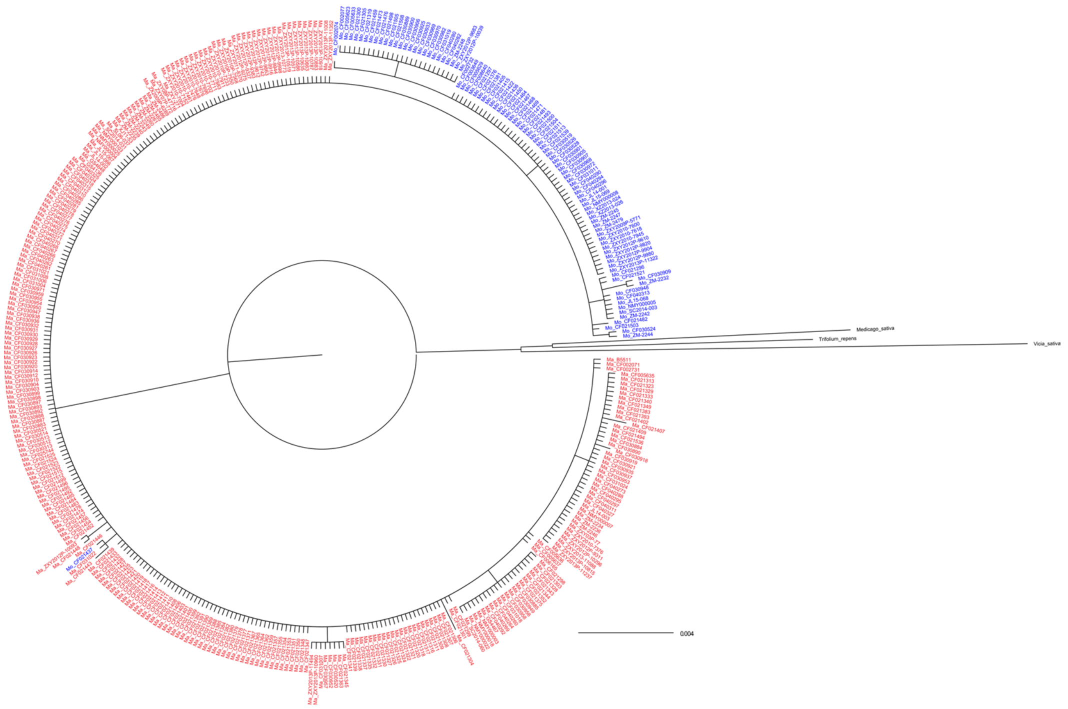

Figure 3. Bayesian tree of *M. albus* and *M. officinalis* with branch lengths, based on ITS sequences. Ma—*M. albus* and Mo—*M. officinalis*. See Supplement Table S3 for accession numbers.

and multiple clustering preferred sampling. The core sets based on different sampling methods have different characteristics and are suitable for different studies. Random sampling can represent the genetic diversity structure of the initial collections and preferred sampling can keep the accessions with special or valuable characteristics in the initial collection[35].

According to multiple clustering random sampling (Table 1), the values of CR% and VR% of *M. albus* did not change significantly as the sampling ratio reached 15%, and then genetic diversity of seed morphology declined smoothly as the sampling proportion increases. For *M. officinalis*, the proper sampling ratio was 25% or 20% based on the values of MD% and CR%, but the nucleotide diversity and haplotype diversity changed steadily until sampling proportion reached 25%. According to multiple clustering preferred sampling, nearly all MD% values are 0 and CR% values are 100%, and the VR% values changed steadily until sampling proportions of *M. albus* and *M. officinalis* reached 15% and 25%, respectively. However, through analysis of H', nucleotide diversity and haplotype diversity, the variation of *M. officinalis* changed steadily from 20% sampling ratio. The core sets that have a good representativeness of the initial collection wouldn't have rapid changes about diversity. To obtain more genetic diversity, the sampling ratios of *M. albus* and *M. officinalis* were determined to be 15% and 25%, respectively.

Overall, the coefficient of variation, genetic diversity index and sequence diversity were increased in the core collections, which was expected because diversity increased after the elimination of similar accessions during the development of the core germplasm sets. Additionally, the genetic diversity of *M. officinalis* is higher than that of *M. albus*, as shown in Table 2, and core collections were listed in Supplementary Table S2. The core collections, which maintained a high level of genetic diversity and were representative of the entire population, can be more efficiently used for breeding and phylogenetic studies than the whole collection.

Discussion

Conservation of plant genetic diversity is essential for present and future human well-being. Over the past few years, there have been many welcome developments in the conservation of forage germplasm resources[36]. As a high-quality forage species, *Melilotus* has many advantages and grows widely in China, and nearly 600 accessions of *Melilotus* were collected in NGBFG, China. In our previous study, we employed 199 accessions of 18 species to analyze genetic diversity[36]. The results indicated that *Melilotus* had high genetic variation among species, and thus, we further studied the genetic diversity and phylogenetic relationships of all *Melilotus* accessions in NGBFG, China. To better protect and utilize these resources, we analyzed the diversity of all accessions in NGBFG based on morphological and molecular data and developed core collections of two species. Morphological and molecular data can be analyzed separately or in combination to determine genetic diversity[37]. In addition, when constructing a core collection, a combination of both phenotypic and genotypic data is thought to be more useful than either one of these individually[38]. Based on seed morphological traits and the ITS and *mat*K sequences of *Melilotus*, we analyzed the genetic diversity of this genus and developed core sets to conserve and utilize germplasm resources efficiently.

Species	Sampling Methods	Sampling Ratio (%)	Evaluation Parameters			
			MD (%)	VD (%)	CR (%)	VR (%)
M. albus	Multiple clustering random sampling	5	0	100	97.6417	162.3625
		10	33.3333	66.6667	97.6417	138.2443
		15	0	100	98.2589	135.9037
		20	0	100	98.6975	132.2542
		25	0	100	98.6975	125.1218
		30	0	66.6667	98.6975	120.1121
	Multiple clustering preferred sampling	5	0	100	100	187.3465
		10	0	100	100	155.2074
		15	0	100	100	141.1182
		20	0	100	100	134.7276
		25	0	100	100	129.4654
		30	0	100	100	124.7173
M. officinalis	Multiple clustering random sampling	5	0	33.3333	72.3255	169.7614
		10	0	66.6667	94.6532	165.3490
		15	0	66.6667	94.6532	137.8036
		20	33.3333	33.3333	94.6532	125.1596
		25	0	0	96.2522	122.9678
		30	0	0	96.2522	118.0300
	Multiple clustering preferred sampling	5	33.3333	66.6667	90.9284	206.2261
		10	0	100	100	180.3112
		15	0	100	100	155.3382
		20	0	66.6667	100	146.3267
		25	0	66.6667	100	136.4903
		30	0	66.6667	100	133.0977

Table 1. Percentage of trait differences between the core collections and the initial collection at five sampling proportions. MD: percentage of significant difference ($\alpha = 0.05$) between each core collection and the initial collection for means of traits, VD: percentage of significant difference ($\alpha = 0.05$) between each core collection and the initial collection for variance of traits, CR%: coincidence rate, VR%: variable rate.

According to Fig. 1 and Supplementary Table S1, the shape and size of seeds showed significant variation among and within species. Seed morphology in *Melilotus* showed a larger Pst parameter than some agronomic traits, such as plant height and dry matter yield[39]. These traits are important for seed establishment and survival[40]. Small-seeded species could produce more seeds for a given amount of energy than large-seeded species; however, large-seeded species, such as *M. italicus* and *M. speciosus*, develop seedlings that can better tolerate the many stresses encountered during establishment[41]. The variations in seed morphology could also reflect the wide range of habitats in *Melilotus*. This information on seed trait variation among accessions could also enhance cultivar development programs that focus on improving seedling survival or seed yield[42]. According to the phylogenetic trees based on the ITS sequences, almost all accessions could be divided by species. The first group, including *M. albus*, *M. suaveolens* was the recently diverged lineages, within the *Melilotus* genus. Additionally, the ITS sequences showed high discrimination in *Melilotus* in this study, while the results revealed that the *mat*K sequences did not perform as well as the ITS sequences. The *mat*K sequences might be more suitable for analyzing relationship at higher taxonomic levels[43], but they can also reflect the variation among and within species to a certain degree[44]. Eighteen species included many subclades, but many accessions within each species showed the same branch lengths in both trees. Although the number of *M. albus* accessions was large, many repetitions were present, because of the frequent exchange of germplasm resources or resubmission of the same accessions. Clarifying the phylogenetic relationship and evaluating the genetic diversity of these accessions will provide a foundation for effective utilization of *Melilotus* accessions in NGBFG.

As the most widely-cultivated species in *Melilotus*, *M. albus* and *M. officinalis* are widely used in forage production and herbal medicine due to the biological activity of their coumarins[39]. Comparing *M. albus* with *M. officinalis*, the seed morphologies are similar (Fig. 1), and in fact, many taxonomic databases, including the USDA PLANTS database, the Integrated Taxonomic Information System, the BugwoodWiki website, and the Catalogue of Life website, have promulgated that the two species are merely conspecific colour morphs that do not merit taxonomic distinction or "accepts" *M. albus* both as a distinct species and as a subspecies of *M. officinalis* due to the similarity of morphological features and growing habits[45]. However, the phylogenetic trees we did in this study (Fig. 3 and Supplementary Fig. S2) with the previous studies[10,23] indicated that *M. albus* and *M. officinalis* have a small genetic distance but are indeed distinct species. Furthermore, we developed core collections of these two species. Genetic parameters and cluster analysis were used to evaluate the efficiency of the development of the core germplasm set[46,47]. In this study, the genetic diversity index, haplotype diversity and nucleotide diversity of the core set were calculated and the core collections were evenly distributed across all clades in phylogenetic trees. Moreover, the sampling rates of *M. albus* and *M. officinalis* were different, which may be due to a difference

Species	Sampling Methods	Sampling Ratio (%)	ITS		*mat*K		H'		
			Haplotype Diversity	Nucleotide Diversity	Haplotype Diversity	Nucleotide Diversity	Length-width Ratio	Seed Circumference	Hundred-seed Weight
M. albus	Multiple clustering random sampling	100	0.414	0.00085	0.547	0.00165	0.0066	0.0063	0.0058
		30	0.42	0.0007	0.68	0.00315	0.0100	0.0096	0.0094
		25	0.411	0.00068	0.631	0.00267	0.0107	0.0109	0.0101
		20	0.419	0.00069	0.62	0.00245	0.0117	0.0119	0.0114
		15	0.444	0.00082	0.684	0.00337	0.0127	0.0128	0.0129
		10	0.538	0.00105	0.664	0.00343	0.0140	0.0136	0.0096
		5	0.714	0.00156	0.705	0.0048	0.1562	0.1592	0.1645
	Multiple clustering preferred sampling	30	0.525	0.00183	0.558	0.00189	0.0106	0.0109	0.0108
		25	0.532	0.00196	0.566	0.00200	0.0111	0.0116	0.0115
		20	0.548	0.00217	0.581	0.00222	0.0120	0.0122	0.0120
		15	0.542	0.00247	0.577	0.00257	0.0127	0.0130	0.0131
		10	0.577	0.00157	0.591	0.00172	0.0141	0.0144	0.0140
		5	0.628	0.00226	0.562	0.00200	0.0160	0.0170	0.0156
M. officinalis	Multiple clustering random sampling	100	0.758	0.00234	0.736	0.00283	0.0104	0.0107	0.0109
		30	0.578	0.00126	0.808	0.00256	0.0145	0.0144	0.0146
		25	0.569	0.00122	0.769	0.00221	0.0146	0.0152	0.0146
		20	0.725	0.00197	0.772	0.00416	0.0153	0.0155	0.0162
		15	0.791	0.0023	0.813	0.00523	0.0164	0.0167	0.0154
		10	0.722	0.00206	0.833	0.00724	0.0179	0.0167	0.0172
		5	0.833	0.00181	0.833	0.00303	0.0178	0.0189	0.0189
	Multiple clustering preferred sampling	30	0.745	0.00380	0.775	0.00385	0.0143	0.0141	0.0135
		25	0.823	0.00440	0.830	0.00428	0.0145	0.0146	0.0140
		20	0.838	0.00515	0.850	0.00499	0.0151	0.0151	0.0148
		15	0.909	0.00348	0.910	0.00338	0.0159	0.0167	0.0160
		10	0.893	0.00350	0.893	0.00350	0.0172	0.0172	0.0174
		5	0.833	0.00350	0.833	0.00350	0.0178	0.0189	0.0178

Table 2. The comparison of the genetic diversity of the total collection *versus* the core sets. H: genetic diversity index calculated using Shannon's information index.

in genetic variation. *Melilotus officinalis* showed higher diversity than *M. albus*, which might be caused by pollination type. *Melilotus albus* is cross-pollinating but self-fertile, while *M. officinalis* is self-incompatible[48].

Core germplasm collections were constructed preliminarily, and additional studies (such as agronomic traits, plant morphology, biochemistry and other molecular marker data) are required to prefect the development of core germplasm collections. Although many rare alleles might not be captured in the core collections, developing core collections could help breeders increase efficiency and utilize genetic resources since cultivar development in *Melilotus* is still in the beginning stage. Besides, the results could also build a foundation for further physiological, genetic and molecular studies in *Melilotus* and provide a reference for future collection and conservation of *Melilotus* and other forages.

Materials and Methods

Plant materials. A total 621 accessions of *Melilotus* were evaluated in the study, and the details of these accessions are presented in Supplementary Table S3. The accessions in NGBFG, China, covered only nine species and most of the accessions belonged to five species, and thus, we added 48 accessions from NPGS, USA, that were studied in the previous study to analyze the phylogenetic structure and genetic diversity in *Melilotus*. To extract DNA, approximately 25 seeds of each accession were polished because of their hardness and then germinated at 24 °C after incubation in a 16-h light/8-h dark cycle. After two weeks, the seedlings were rinsed by distilled water, collected separately, frozen in liquid nitrogen and maintained at −80 °C until extracted.

Seed morphology. Five characters of seeds were measured, including length, width, width-to-length ratio, circumference and 100-seed weight. We selected 100 seeds of each accession at random and measure their morphology using an analytical balance and WinSEEDLE, an image analysis system for morphological and disease analysis of seeds and needles.

DNA extraction, amplification, and sequencing. Total genomic DNA was extracted from whole seedling material according to the SDS (sodium dodecyl sulfate) method[49]. The target DNA fragments, the internal transcribed spacer (ITS) and chloroplast locus *mat*K, were amplified and sequenced[50,51]. Amplification was performed by polymerase chain reactions (PCR) in 25-μL mixtures containing 12.25 μL of 2× reaction mix, 2 μL of each primer (1 μmol/mL), 2 μL of template genomic DNA (50 ng/μL), 0.25 μL of Golden DNA polymerase and 6.5 μL of deionized water. The primers and details of amplification programs were listed in Supplementary Table S4. Successful PCR products were sent to Shanghai Shenggong Biotechnological Ltd. (Shanghai, China) for sequencing.

Alignment and diversity analysis. Both ends of the DNA sequences were trimmed to remove unalignable sequences upstream and downstream of the homologous sites by the Contig Express module of Vector NTI Suite 6.0 (InforMax, Inc) and aligned by DNAMAN 7.0[52,53]. The haplotype diversity and nucleotide diversity were computed by DnaSP 6.11[54]. The phylogenetic trees were drawn by ClustalW of MEGA 6.0 and MrBayes 3.2 software. The Bayesian method was adopted with the default settings and the GTR model with gamma-distributed rate variation across sites and a proportion of invariable sites (nst, 6; rates, invgamma)[55] and operational generation number and sampling frequency were set to 100000000 and 100000, with *Medicago sativa*, *Trifolium repens* and *Vicia sativa* as outgroups. The morphological traits were analyzed using the statistical software package SPSS v16.0[37]. The coefficient of variation, phenotypic differentiation and Shannon' genetic diversity index (H') were calculated to analyze seed morphological diversity. The phenotypic differentiation coefficient (Pst) was calculated as follows: $Pst = (\sigma^2_{t/s})/(\sigma^2_{t/s} + \sigma^2_s)$, where $\sigma^2_{t/s}$ is the variance portion among populations and σ^2_s is the variance portion within populations[56]. Shannon's diversity index was calculated as follows: $H' = -\sum pi \text{ Ln } pi$, where *pi* is the proportion of each phenotypic trait[57].

Development of core collections. We used QGAStation 2.0, a software for classical quantitative genetics, to construct a core set according to the seed morphology. The strategy for constructing core collections adopted the least distance stepwise sampling based on genotypic values[58], and Hu *et al.* (2000) suggested that standardized Euclidean distance combined with nearest distance method was an appropriate genetic distance for constructing core collections in this strategy[35]. We tried two sampling methods, multiple clustering random sampling and multiple clustering preferred sampling, to determine the appropriate sampling method and proportions[35]. Multiple clustering random sampling: one accession from each subgroup with two accessions at the lowest level of sorting is randomly selected. If there is only one accession in a subgroup, it is directly sampled for the next cluster. Multiple clustering preferred sampling: accessions with maximum or minimum values of traits are preferred to select from each subgroup at the lowest level of sorting. Both accessions are selected if two accessions in a subgroup have maximum or minimum values of the traits. The other procedures are similar to the random sampling strategy.

Six sampling proportions were chosen in the study, which were 5%, 10%, 15%, 20%, 25% and 30%. We calculated four parameters to evaluate the representation of the core germplasm at different sampling rates[58]: mean difference percentage (MD%), variance difference percentage (VD%), coincidence rate of range (CR%) and changeable rate of coefficient of variation (VR%). Additionally, the Shannon' genetic diversity index of seed morphology and the haplotype diversity and nucleotide diversity of sequences were calculated to assess the genetic diversity of the core collections. According to the genetic diversity comparison of these core collections, we could determine the best sampling proportion, which was considered to be representative while maintaining a high level of genetic diversity.

References

1. Smith, W. K. & Gorz, H. J. Sweetclover Improvement. *Adv. Agron.* **17**, 163–231, https://doi.org/10.1016/S0065-2113(08)60414-9 (1965).
2. Stevenson, G. A. An agronomic and taxonomic review of the genus *Melilotus* Mill. *Can. J. Plant Sci.* **49**, 1–20, https://doi.org/10.4141/cjps69-001 (1969).
3. Barnes, D. K., Sheaffer, C. C., Heath, M. E., Barnes, R. F. & Metcalfe, D. S. Forages: the science of grassland agriculture. *J. Range Manage.* **38**, 382, https://doi.org/10.2307/3894695 (1952).
4. Rogers, M. E. *et al.* Diversity in the genus *Melilotus* for tolerance to salinity and waterlogging. *Plant & Soil* **304**, 89–101, https://doi.org/10.1007/s11104-007-9523-y (2008).
5. Sherif, E. A. A. *Melilotus indicus* (L.) All., a salt-tolerant wild leguminous herb with high potential for use as a forage crop in salt-affected soils. *Flora* **204**, 737–746, https://doi.org/10.1016/j.flora.2008.10.004 (2009).
6. Stickler, F. C. & Johnson, I. J. Dry Matter and Nitrogen Production of Legumes and Legume Associations in the Fall of the Seeding Year1. *Agron. J.* **51**, 135–137, https://doi.org/10.2134/agronj1959.00021962005100030004x (1959).
7. Campbell, C. A., Bowren, K. E., Schnitzer, M., Zentner, R. P. & Townleysmith, L. Effect of crop rotations and fertilization on soil organic matter and some biochemical properties of a thick Black Chernozem. *Can. J. Soil Sci.* **71**, 377–387, https://doi.org/10.4141/cjss91-036 (1991).
8. Rusterholz, H. P. & Erhardt, A. Effects of elevated CO_2 on flowering phenology and nectar production of nectar plants important for butterflies of calcareous grasslands. *Oecologia* **113**, 341–349, https://doi.org/10.1007/s004420050385 (1998).
9. Barot, K. P., Jain, S. V., Kremer, L., Singh, S. & Ghate, M. D. Recent advances and therapeutic journey of coumarins: current status and perspectives. *Med. Chem. Res.* **24**, 2771–2798 (2015).
10. Wu, F. *et al.* Analysis of genetic diversity and population structure in accessions of the genus *Melilotus*. *Ind. Crop. Prod.* **85**, 84–92, https://doi.org/10.1016/j.indcrop.2016.02.055 (2016).
11. Nair, R. *et al.* Variation in coumarin content of Melilotus species grown in South Australia. *New Zeal. J. Agr. Res.* **53**, 201–213, https://doi.org/10.1080/00288233.2010.495743 (2010).
12. Su, W. *et al.* Genome-wide assessment of population structure and genetic diversity and development of a core germplasm set for sweet potato based on specific length amplified fragment (SLAF) sequencing. *PloS One* **12**, e0172066, https://doi.org/10.1371/journal.pone.0172066 (2017).
13. Roy, C. D. *et al.* Analysis of genetic diversity and population structure of rice germplasm from north-eastern region of India and development of a core germplasm set. *PloS One* **9**, e113094, https://doi.org/10.1371/journal.pone.0113094 (2014).
14. Kim, S. K., Nair, R. M., Lee, J. & Lee, S. H. Genomic resources in mungbean for future breeding programs. *Front. Plant Sci.* **6**, 626, https://doi.org/10.3389/fpls.2015.00626 (2015).
15. Rao, V. R. & Hodgkin, T. Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell Tiss. Org.* **68**, 1–19, https://doi.org/10.1023/A:1013359015812 (2002).
16. Wang, J. *et al.* A Strategy for Finding the Optimal Scale of Plant Core Collection Based on Monte Carlo Simulation. *The Scientific World J.,2014,(2014-1-20)* **2014**, 503473, https://doi.org/10.1155/2014/503473 (2014).

17. Ebana, K., Kojima, Y., Fukuoka, S., Nagamine, T. & Kawase, M. Development of mini core collection of Japanese rice landrace. *Breeding Sci.* **58**, 281–291, https://doi.org/10.1270/jsbbs.58.281 (2008).
18. Yan, W. G. *et al*. Development and Evaluation of a Core Subset of the USDA Rice Germplasm Collection. *Crop Sci.* **47**, 869–878, https://doi.org/10.2135/cropsci2006.07.0444 (2007).
19. Holbrook, C. C. & Dong, W. Development and evaluation of a mini core collection for the U.S. peanut germplasm collection. *Crop Sci.* **45**, 1540–1544, https://doi.org/10.2135/cropsci2004.0368 (2005).
20. Li, Y., Shi, Y., Cao, Y. & Wang, T. Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. *Genet. Resour. Crop Ev.* **51**, 845–852, https://doi.org/10.1007/s10722-005-8313-8 (2005).
21. Díez, C. M., Imperato, A., Rallo, L., Barranco, D. & Trujillo, I. Worldwide Core Collection of Olive Cultivars Based on Simple Sequence Repeat and Morphological Markers. *Crop Sci.* **52**, 211–221, https://doi.org/10.2135/cropsci2011.02.0110 (2012).
22. Liang, W. *et al*. Genetic Diversity, Population Structure and Construction of a Core Collection of Apple Cultivars from Italian Germplasm. *Plant Mol. Biol. Rep.* **33**, 458–473, https://doi.org/10.1007/s11105-014-0754-9 (2015).
23. Zhang, H. *et al*. Genetic variation and diversity in 199 *Melilotus* accessions based on a combination of 5 DNA sequences. *PloS ONE* **13**, https://doi.org/10.1371/journal.pone.0194172 (2018).
24. Grillo, O., Mattana, E., Venora, G. & Bacchetta, G. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Sci. Technol.* **38**(455–476), 422, https://doi.org/10.15258/sst.2010.38.2.19 (2010).
25. Balkaya, A., Yanmaz, R. & Özer, M. Evaluation of variation in seed characters of Turkish winter squash (*Cucurbita maxima*) populations. *New Zeal. J. Exp. Agr.* **37**, 167–178, https://doi.org/10.1080/01140670909510262 (2009).
26. Alghamdi, F. A. & Alzahrani, R. M. Seed morphology of some species of Tephrosia Pers. (Fabaceae) from Saudi Arabia identification of species and systematic significance. *Am. J. Plant Sci.* **2**, 484–495 (2010).
27. Moody, M. L., Hufford Lsoltis, D. E. & Soltis, P. S. Phylogenetic relationships of Loasaceae subfamily Gronovioideae inferred from matK and ITS sequence data. *Am. J. Bot.* **88**, 326–336, https://doi.org/10.2307/2657022 (2001).
28. Lü, N., Yamane, K. & Ohnishi, O. Genetic diversity of cultivated and wild radish and phylogenetic relationships among Raphanus and Brassica species revealed by the analysis of *trn*K/*mat*K sequence. *Jap. J. Breeding* **58**, 15–22, https://doi.org/10.1270/jsbbs.58.15 (2008).
29. Zhang, W., Kan, S. L., Zhao, H., Li, Z. Y. & Wang, X. Q. Molecular phylogeny of tribe Theeae (Theaceae s.s.) and its implications for generic delimitation. *PloS One* **9**, e98133, https://doi.org/10.1371/journal.pone.0098133 (2014).
30. Steiner, J. J. Molecular phylogenetics of the clover genus (Trifolium–Leguminosae). *Molecular Phylogenetics & Evolution* **39**, 688–705, https://doi.org/10.1016/j.ympev.2006.01.004 (2006).
31. Lavin, M., Herendeen, P. S., Wojciechowski, M. F. & Linder, P. Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. *Systematic Biol.* **54**, 575–594, https://doi.org/10.1080/10635150590947131 (2005).
32. Di, H. *et al*. Interspecific Phylogenic Relationships within Genus *Melilotus* Based on Nuclear and Chloroplast DNA. *PloS One* **10**, e0132596, https://doi.org/10.1371/journal.pone.0132596 (2014).
33. Smith, W. K. Viability of Interspecific Hybrids in *Melilotus*. *Genet.* **39**, 266–279, https://doi.org/10.2307/2405786 (1954).
34. Wang, L. L., Wang, E. T., Liu, J., Li, Y. & Chen, W. X. Endophytic occupation of root nodules and roots of *Melilotus dentatus* by Agrobacterium tumefaciens. *Microbial Ecol.* **52**, 436–443, https://doi.org/10.1007/s00248-006-9116-y (2006).
35. Hu, J., Zhu, J. & Xu, H. M. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor. Appl. Genet.* **101**, 264–268, https://doi.org/10.1007/s001220051478 (2000).
36. Bai, C. J. *et al*. Technical challenges in evaluating southern China's forage germplasm resources. *Tropical Grasslands-Forrajes Tropicales* **1**, 184, https://doi.org/10.17138/TGFT(1)184-191 (2013).
37. Putcha, V. Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS. *J. Appl. Stat.* **42**, 2291–2291, https://doi.org/10.1111/j.1467-985X.2007.00521_10.x (2010).
38. Li, X. *et al*. Genotypic and phenotypic characterization of genetic differentiation and diversity in the USDA rice mini-core collection. *Genetica* **138**, 1221–1230, https://doi.org/10.1007/s10709-010-9521-5 (2010).
39. Zhang, J. *et al*. Coumarin Content, Morphological Variation, and Molecular Phylogenetics of *Melilotus*. *Molecules* **23**, https://doi.org/10.3390/molecules23040810 (2018).
40. Aarssen, L. W. & Jordan, C. Y. Between-species patterns of covariation in plant size, seed size and fecundity in monocarpic herbs. *Ecoscience* **8**, 471–477, https://doi.org/10.1080/11956860.2001.11682677 (2001).
41. Henery, M. L. & Westoby, M. Seed mass and seed nutrient content as predictors of seed output variation between species. *Oikos* **92**, 479–490, https://doi.org/10.1034/j.1600-0706.2001.920309.x (2001).
42. Tíscar Oliver, P. A., Lucas Borja, M. E. & Bravo, F. Seed mass variation, germination time and seedling performance in a population of Pinus nigra subsp. salzamannii. *Forest Syst.* **19**, 344–353, https://doi.org/10.5424/fs/2010193-9094 (2010).
43. Johnson, L. A. & Soltis, D. E. matK DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. *Syst. Bot* **19**, 143–156, https://doi.org/10.2307/2419718 (1994).
44. Bai, W. N. & Zhang, D. Y. Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. *New Phytologist* **188**, 892–901, https://doi.org/10.1111/j.1469-8137.2010.03407.x (2010).
45. Darbyshire, S. & Small, E. Are *Melilotus albus* and *M. officinalis* conspecific? *Genet. Resour. Crop Ev.* 1–10, https://doi.org/10.1007/s10722-018-0627-4 (2018).
46. He, S., Wang, Y., Volis, S., Li, D. & Yi, T. Genetic Diversity and Population Structure: Implications for Conservation of Wild Soybean (Glycine soja Sieb. et Zucc) Based on Nuclear and Chloroplast Microsatellite Variation. *Int. J. Mol. Sci.* **13**, 12608, https://doi.org/10.3390/ijms131012608 (2012).
47. Jansen, J. & Van Hintum, T. Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Tag. theor. Appl. Genet. theoretische Und Angewandte Genetik* **114**, 421, https://doi.org/10.1007/s00122-006-0433-9 (2007).
48. Turkington, R. A., Cavers, P. B. & Rempel, E. The biology of Canadian weeds. 29. *Melilotus alba* Desr. and *M. officinalis* (L.) Lam. *Revue Canadienne De Phytotechnie* **58**, 523–537, https://doi.org/10.4141/cjps78-078 (1978).
49. Zhao, X. C. & Sharp, P. J. An improved 1-D SDS-PAGE method for the identification of three bread wheat 'waxy' proteins. *J. Cereal Sci.* **23**, 191–193, https://doi.org/10.1006/jcrs.1996.0019 (1996).
50. Chiou, S. J., Yen, J. H., Fang, C. L., Chen, H. L. & Lin, T. Y. Authentication of Medicinal Herbs using PCR-Amplified ITS2 with Specific Primers. *Planta Medica* **73**, 1421–1426, https://doi.org/10.1055/S-2007-990227 (2007).
51. Bafeel, S. O. *et al*. Comparative evaluation of PCR success with universal primers of maturase K (*mat*K) and ribulose-1, 5-bisphosphate carboxylase oxygenase large subunit *(rbc*L) for barcoding of some arid plants. *Plant Omics*, 195–198, https://doi.org/10.1111/j.1438-8677.2010.00434.x (2011).
52. Lu, G. & Moriyama, E. N. Lu, G. & Moriyama, E. N. Vector NTI, a balanced all-in-one sequence analysis suite. *Brief. Bioinform*. 5, 378–388. *Briefings in Bioinformatics* **5**, 378–388, https://doi.org/10.1093/bib/5.4.378 (2005).
53. Brady, S. G., Schultz, T. R., Fisher, B. L. & Ward, P. S. Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proc Natl Acad Sci USA* **103**, 18172–18177, https://doi.org/10.1073/pnas.0605858103 (2006).
54. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452, https://doi.org/10.1093/bioinformatics/btp187 (2009).

55. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755, https://doi.org/10.1093/bioinformatics/17.8.754 (2001).
56. Zongyu, Z. *et al*. Phenotype- and SSR-Based Estimates of Genetic Variation between and within Two Important *Elymus* Species in Western and Northern China. *Genes.* **9**(3), 147-, https://doi.org/10.3390/genes9030147 (2018).
57. Zhang, M. Q. *et al*. A study on genetic diversity of reproductive characters in *Elymus nutans* germplasm resources. *Acta Prataclt. Sin.* **20**, 182–191, https://doi.org/10.1093/mp/ssq070 (2011).
58. Wang, J. C., Hu, J., Xu, H. M. & Zhang, S. A strategy on constructing core collections by least distance stepwise sampling. *Theor Appl Genet.* **115**(1), 1–8, https://doi.org/10.1007/s00122-007-0533-1 (2007).

Acknowledgements

This work was supported by the National Basic Research Program (973) of China (2014CB138704), the National Natural Science Foundation of China (31572453), the Program for Changjiang Scholars and the Innovative Research Team in Chinese Universities (IRT_17R50), the Open Project Program of the State Key Laboratory of Grassland Agroecosystems (SKLGAE201702), and the 111 project (B12002). Additionally, we thank the NGBFG and NPGS for providing the experimental materials used in our study.

Author Contributions

Conceptualization, Jiyu Zhang; Data curation, Hongxiang Zhang, Rong Bai and Wenli Guo; Formal analysis, Rong Bai and Jinxing Ma; Funding acquisition, Jiyu Zhang; Investigation, Wenli Guo, Zhuanzhuan Yan and Yufei Zhang; Methodology, Fan Wu and Yufei Zhang; Project administration, Jiyu Zhang; Software, Hongxiang Zhang, Zhuanzhuan Yan and Qi Yan; Supervision, Jinxing Ma; Visualization, Qi Yan; Writing – original draft, Hongxiang Zhang and Rong Bai; Writing – review & editing, Fan Wu and Jiyu Zhang.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49355-y.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.