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

Hongxiang Zhang¹,², Rong Bai¹, Fan Wu¹, Wenli Guo¹, Zhuanzhuan Yan¹, Qi Yan¹, Yufei Zhang¹, Jinxing Ma³ & Jiyu Zhang¹

*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’s genetic diversity index (H’) and phenotypic differentiation (Pst) were calculated to better describe the genetic diversity. The ITS and matK sequences were used to construct phylogenetic trees and study the genetic relationships within genus *Melilotus*. 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*. In comparison with most other forages, *Melilotus* has the advantages of tolerating extreme environmental conditions and providing high seed yields. The nitrogen fixation rate of *Melilotus* is superior to those of other legumes, and it is beneficial in crop rotations. Additionally, *Melilotus* can be used as a crop fertilizer 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. During the past few years, *Melilotus*, as a good leguminous forage, has received much attention. 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. Abundant genetic resources have great potential to provide novel beneficial genes.

During the last 3–4 decades, major advances have been made in conserving these resources. 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. 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

¹State 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. ²State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, P.R. China. ³National 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)
been proposed for better management and use of the collections available in gene banks. 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. 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. 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. Seed morphology and the sequences of ITS and *matK* 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. More importantly, seed morphological traits can be utilized for species identification as well as selection criteria in crop improvement programs. The nuclear DNA ITS and chloroplast DNA *matK* have been widely applied in studies of inferring phylogenetic relationships at lower taxonomic levels and have been successfully used to analyze plant systematics. The previous studies in Fabaceae indicated that the rate and pattern of ITS sequence mutation are appropriate for resolving relationships among species and genera, as well as revealed that *matK* sequence can be used in phylogenetic analyses to successfully resolve relationships even at the species level. Additionally, these sequences showed high stability and discrimination in *Melilotus*. 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 cm, 6.564 cm and 0.365 g, respectively. According to Supplementary Table S1, an analysis of variance indicated significant 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 morphology 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 *matK* sequences also performed well, with a
the previous study, in which 18 species formed two groups. Ten species, which were abbreviated to represent 18 species: Ma—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.

Figure 2. Bayesian tree of 18 species in Melilotus with branch lengths, based on ITS sequences. The abbreviations represent 18 species: Ma—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.

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. We tried two different sampling methods, multiple clustering random sampling.
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.

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. 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. 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. 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. Based on seed morphological traits and the ITS and *matK* sequences of *Melilotus*, we analyzed the genetic diversity of this genus and developed core sets to conserve and utilize germplasm resources efficiently.
Melilotus

The ITS sequences showed high discrimination in mat sequences did not perform as well as the ITS sequences. The ing relationship at higher taxonomic levels43, but they can also reflect the variation among and within species to

duction and herbal medicine due to the biological activity of their coumarins 39. Comparing the core germplasm set46,47. In this study, the genetic diversity index, haplotype diversity and nucleotide diversity 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 a certain degree44. Eighteen species included many subclades, but many accessions within each species showed

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 yield39. These traits are important for seed establishment and survival40. 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 establishment41. 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 yield42. 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 matK sequences did not perform as well as the ITS sequences. The matK sequences might be more suitable for analyzing relationship at higher taxonomic levels43, but they can also reflect the variation among and within species to a certain degree44. 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 coumarins39. 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 habits45. However, the phylogenetic trees we did in this study (Fig. 3 and Supplementary Fig. S2) with the previous studies10,23 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 set46,47. 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 (%) | Evaluation Parameters |
|---------|------------------|--------------------|-----------------------|
|         |                  | MD (%)          | VD (%)         | CR (%) | VR (%)|
| M. albus | Multiple clustering | 5 0 100 97.6417 162.3625 | random sampling   | 10 33.3333 66.6667 97.6417 138.2443 |
|         | preferred sampling | 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 |
| M. officinalis | Multiple clustering | 5 0 100 100 187.3465 | random sampling   | 10 0 100 100 155.2074 |
|         | preferred sampling | 15 0 100 100 141.1182 |                     | 20 0 100 100 134.7276   |
|         |                  | 25 0 100 100 129.4654 |                     | 30 0 100 100 124.7173  |
|         |                  | 5 0 33.3333 72.3255 169.7614 | random sampling   | 10 0 100 100 180.3112 |
|         | preferred sampling | 15 0 100 100 124.7173 |                     | 20 33.3333 33.3333 94.6532 165.3490 |
|         |                  | 25 0 100 100 124.7173 |                     | 30 0 66.6667 96.2522 118.0300   |
|         |                  | 5 33.3333 66.6667 90.9284 206.2261 | preferred sampling | 10 0 100 100 155.3382 |
|         |                  | 15 0 100 100 137.8036 |                     | 20 33.3333 33.3333 94.6532 125.1596 |
|         |                  | 25 0 100 100 129.4654 |                     | 30 0 66.6667 96.2522 118.0300   |
|         |                  | 5 33.3333 66.6667 90.9284 206.2261 | preferred sampling | 10 0 100 100 180.3112 |
|         |                  | 15 0 100 100 155.3382 |                     | 20 33.3333 33.3333 94.6532 125.1596 |
|         |                  | 25 0 100 100 129.4654 |                     | 30 0 66.6667 96.2522 118.0300   |

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

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 morphological material according to the SDS (sodium dodecyl sulfate) method. The target DNA fragments, the internal transcribed spacer (ITS) and chloroplast locus *matK*, were amplified and sequenced. 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/µL), 2 µL of template genomic DNA (50 ng/µL), 0.25 µL of Golden DNA polymerase

| Species | Sampling Methods | Sampling Ratio (%) | Haplotypes Diversity | Nucleotide Diversity | Haplotypes Diversity | Nucleotide Diversity | Length-width Ratio | Seed Circumference | Hundred-seed Weight |
|---------|------------------|-------------------|---------------------|---------------------|---------------------|---------------------|-------------------|-------------------|--------------------|
| *M. albus* | Multiple clustering random sampling | 100 | 0.414 | 0.00858 | 0.547 | 0.00165 | 0.0066 | 0.0063 | 0.0058 |
| | | 30 | 0.42 | 0.00078 | 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 | 0.777 | 0.00416 | 0.0153 | 0.0155 | 0.0162 |
| | | 15 | 0.791 | 0.0022 | 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.

---
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 6.11.24. 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 v16.0.57. 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 = (σ^2_uk)/(σ^2_uk + σ^2_uq), where σ^2_uq is the variance portion among populations and σ^2_uk is the variance portion within populations56. Shannon’s diversity index was calculated as follows: H' = - Σ p_i ln p_i, where p_i is the proportion of each phenotypic trait57.

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 values58, 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 strategy59. We tried two sampling methods, multiple clustering random sampling and multiple clustering preferred sampling, to determine the appropriate sampling method and proportions30,31. 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 rates55, 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. Sweet clover improvement. Adv. Agron. 17, 163–231, https://doi.org/10.1016/S0065-2113(08)60144-9 (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/10.2137/3804695 (1982).
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. Stuckler, F. C. & Johnson, I. J. Dry Matter and Nitrogen Production of Legumes and Legume Associations in the Fall of the Seeding Year. Agron. J. 51, 135–137, https://doi.org/10.2134/agronj1959.00021962005100003004xs (1959).
7. Campbell, C. A., Bowren, K. E., Schnitzius, M., Zentner, R. P. & Townley-Smith, 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/10.1139/cjs91-036 (1991).
8. Rusterholz, H. P. & Erhardt, A. Effects of elevated CO2 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. & Ghatte, 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 cowpea species and 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 diversity 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/10.1007/s11240-001-9582-7 (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).
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 Pratacult. Sin. 20, 182–191, https://doi.org/10.1093/mp/qsq070 (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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019