Journal of Genetic Engineering and Biotechnology 23 (2025) 100491

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

Genetic diversity and population structure analysis in *Asparagus officinalis*

Gurleen Sidhu [a,*], Travis Banks [b], David Wolyn [a]

[a] Department of Plant Agriculture, University of Guelph, Guelph, Ontario, Canada
[b] Vineland Research and Innovation Centre, Vineland, Ontario, Canada

ARTICLE INFO

Keywords:
Genetic diversity
Asparagus
SNP markers
Genotyping-by-sequencing

ABSTRACT

Asparagus cultivars grown worldwide are thought to have originated from a limited genetic base, however, selection has resulted in variation for climatic adaptation and other traits. Understanding genetic diversity of the crop is important to guide breeding decisions. The objectives of this research were to identify SNPs among 64 cultivated lines from different geographic areas and assess genetic variation, population structure and linkage disequilibrium. More than 55,000 SNPs were identified by GBS and subjected to filtration for minor allele frequency and missing data, resulting in 12,886 markers for all subsequent analysis. Markers exhibited a wide range of Expected Heterozygosity (He), Polymorphic Information Content (PIC) and Observed Heterozygosity (Ho) with mean values of 0.370, 0.310, and 0.450 respectively. Population STRUCTURE analysis indicated that the 64 lines were grouped into two, four, seven, and nine subpopulations. For K = 4, four distinct groups were defined: (1) New Zealand, New Jersey, France, and California; (2) Canada; (3) China, The Netherlands, and Germany; and (4) England, Denmark, Spain, Turkey, and India. The results were further confirmed by PCA, and a phylogenetic tree. LD declined rapidly with an increase in physical distance. A considerable amount of genetic diversity was observed, despite previous suggestions that asparagus cultivars may have originated from one open-pollinated population.

1. Introduction

Asparagus (*Asparagus officinalis* L.) is an herbaceous perennial vegetable where the emerging stems, with abundant nutritional and medicinal properties, are the harvestable product [43]. A dioecious species with high levels of cross-pollination, asparagus is native to most of Europe, northern Africa and western Asia and was grown in the ancient Roman Empire [23]. Cultivation became prevalent in the 16th and 17th centuries in France, Germany, England, and The Netherlands, where specific populations were recognized. Asparagus is grown in over 60 countries and on most continents.

Violet Dutch was the first identified asparagus population, from which most cultivars today have been derived [6]. Breeding Violet Dutch resulted in two populations, Argenteuil in France and Braunschweiger in Germany which replaced old populations and landraces. Cultivars from the USA, Taiwan, France, and Italy were derived mostly from Argenteuil and those from Germany and The Netherlands were derived from Braunschweiger [31].

Based on the history of asparagus production, two different origins are suggested. Cultivars grown in United States, France and Italy were derived from Argenteuil population whereas those grown in Germany and the Netherlands were selected from the Braunschweiger population [23]. Even though both populations were derived from a single source population, they were kept separately for many years.

Studies have shown the formation of separated groups based on geographic origin (Argenteuil and Braunschweiger). Cultivars have been grouped based on geographic origin. Assessing 26 cultivars from different origins using morphological and isoenzymatic markers separated germplasm into to two groups, related to ancestral populations, Argenteuil and Braunschweiger [15]). An assessment of genetic diversity using EST-SSR markers in 30 asparagus cultivars resulted in two groupings, one included cultivars from America, Spain, Italy and Germany, and the other cluster comprised cultivars from The Netherlands, Germany, Spain, and New Zealand. Data suggested asparagus grown in Europe and America were derived from different germplasm pools with some cultivars having genes from both population sources [7].

Abbreviations: He, Expected Heterozygosity; PIC, Polymorphic Information Content; Ho, Observed Heterozygosity; TASSEL, Trait Analysis by Association Evolution and Linkage; $F_{st:}$, fixation index; GAPIT, Genomic Association and Prediction Integrated Tool; AMOVA, Analysis of Molecular Variance; IBS, Identity by State; EMMA, Efficient Mixed Model Association; IBD, Identity by Descent; LD, Linkage Disequilibrium; ISSR, Inter-Simple Sequence Repeats.

* Corresponding author.
E-mail address: gsidhu04@uoguelph.ca (G. Sidhu).

https://doi.org/10.1016/j.jgeb.2025.100491
Received 6 December 2024; Received in revised form 21 March 2025; Accepted 7 April 2025

Consequently, North American and European breeders generally selected within their respective germplasm base, although some cultivars with complex origins had germplasm from both groups. Recently, cluster analysis identified two germplasm groups that were not consistent with country of origin [9]; cultivars from North America, Netherlands and China were in both groups. Other studies could also not group germplasm based on country of origin [3,32]. In another asparagus study, using Nei's unbiased genetic distance technique, 24 cultivars were divided into two major groupings, with one cultivar in Cluster I and the remaining 23 cultivars in Cluster II. Both the clusters failed to divide germplasm on the basis of country of origin[1]. Likewise, Li et al. [25] could not distinguish 10 asparagus cultivars using SSR markers based on geographic origin. Diversity studies including cultivated and wild germplasm, and species of different ploidy suggested a narrow genetic base of cultivated diploid asparagus, high diversity of wild accessions, and the potential to increase diploid diversity by introgression from hexaploid species [14,2], Van Cutsem et al., 2003, [32]. Another comprehensive genome-wide diversity analysis was carried out in 2023 by Sala et al. among 378 genotypes of asparagus. A range of genotypes were included in this collection, including doubled haploids, commercial hybrids, interspecific lines, and wild relatives of garden asparagus (*Asparagus officinalis*). The results showed that the cultivated and wild genotypes of *A. officinalis* had a narrow genetic base, in comparison to the diversity found in wild species and interspecific crosses [40].

Although asparagus is thought to have a narrow genetic base, selection for different characteristics and environments has resulted in phenotypic variation among cultivars. The crop is commonly bred for green vs. white production, where harvesting occurs when spears are above or below ground, respectively. Geographically, white production occurs predominantly in Europe, while green is common in most other jurisdictions [34]. Climatic adaptation varies, with different cultivars bred for the cold winters of temperate areas, the hot, dry summers, and cool, wet winters, of the Mediterranean region, as well as tropical and subtropical environments, where temperatures remain warm throughout the year. Thus, germplasm divergence may have occurred based on climatic adaptation.

In asparagus, the use of DNA-based molecular markers to study genetic diversity is limited. Most previous studies were conducted using molecular markers such as RAPD (Random Amplified Polymorphic DNA) [20,41], SCAR (Sequence Characterized Amplified Region) [20], EST-SSR (Expressed Sequence Tag-derived), SSR (Simple Sequence Repeat Markers) (Cuesta-Marcos et al., 2008; [22], SRAP (Sequence Related Amplified Polymorphism)[3] and AFLP (Amplified Fragment Length Polymorphism) [38]. Such systems provide a limited number of markers and restrict population size. SNP genotyping with genotyping-by-sequencing methods (reference) can allowed thorough diversity analysis in a large number of individuals.

Although significant genetic variation may occur in different species of the *Asparagus* genus, often with polyploid genomes, introgression into cultivated germplasm can be slow and difficult. Consequently, breeding and selection with cultivated germplasm is most often used to achieve rapid results. With limited cultivated germplasm assessed in previous diversity studies, often with restricted numbers of legacy molecular markers, the objectives of this research were to: 1) analyze the genetic variability among 6 different lines of asparagus using SNPs; and 2) determine the genetic relationships among globally distributed asparagus germplasm based on country of origin and commercial use.

2. Materials and methods

2.1. Plant material

Sixty-four asparagus germplasm lines from 13 geographical areas were studied (Additional File 1: Table S1 and Table S2). Twelve were obtained from the US National Plant Germplasm (NPGS, https://www.ars-grin.gov/npgs/) and the remaining fifty-two were obtained from breeding programs which includes nine lines from New Jersey USA, 15 from Ontario Canada, three from Denmark, 12 from The Netherlands, two from Spain, seven from Germany, one from France, two from England, six from New Zealand, one from China, one from India, two from Turkey, and three from California USA. Of these, 25 lines represented white asparagus and the remaining 39 represented green asparagus lines.

2.2. DNA isolation

Young cladophylls from each of five plants were bulked for each line and stored in 2 ml tubes for genomic DNA extraction. The tubes were stored at −20 °C, then tissue samples were freeze-dried (−50 °C, vacuum 0.5 mm Hg; 24 h). Genomic DNA was extracted using commercially available NucleoSpin Plant II kits (Machery-Nagel, Bethlehem, PA, USA) according to manufacturer's instruction. The quantity of extracted DNA was determined by a Qubit™ 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA),and quality assessed with a Nanodrop® spectrophotometer (ND-1000 v.3.5.2; NanoDrop Technologies, Inc., Wilmington, DE, USA) and by agarose gel electrophoresis. DNA samples were then diluted to 10 ng/μl.

Genotyping-by-Sequencing was conducted by Université Laval's Plate-forme D'analyses Génomiques (Université Laval, Québec, Canada). The enzymes MspI and PstI were used to generate the library[30]. Sequencing was completed on an Ion Proton sequencing machine (Life Technologies, Bleiswijk, The Netherlands). A total of 99,415,942 reads of length 50–135 base pairs were generated. All bioinformatics steps were performed with the FastGBS pipeline [44]. Sequence libraries were de-multiplexed with sabre and sequence adapters removed with cutadaptv 1.4.1 [29]. Sample reads were aligned onto the asparagus reference genome (GenBank assembly accession: GCA_001876935.1) with the algorithm BWA mem [24]. Transformation of sam to bam files, and sorting and indexing of bam files were performed with Samtools[46]. Variants (SNPs and short Indels) were called using Platypus [39]. Finally, 56,533 variants that passed platypus filters were kept for subsequent analysis. Minor allele filtering was performed with Trait Analysis by Association Evolution and Linkage (TASSEL) software[5] with the help of filter site. Of the 56,533 SNPs identified initially, 12,886 were used after eliminating those with a minor allele frequency < 0.05 and more than 80 % missing data.

2.3. Data analysis

Structure Analysis, Principal Component Analysis, and Kinship Matrix: The model based genetic clustering algorithm in the STRUCTURE v2.3.3 software package was used to determine the population structure of all lines using SNP markers distributed on all 10 asparagus chromosomes [35]. Structure was determined using the default admixture model based on K-values ranging from 1 to 10. For each K value three independent runs were performed with a burn in length of 50,000 and 100,000 iterations. Subsequently, the optimum number of clusters was determined following the method described by Evanno et al. [13] and the results obtained were visualized using STRUCTURE Harvester [11]. The fixation index (F_{st}) values for nine subpopulations were determined using STRUCTURE software. Allele frequency divergence (net nucleotide distance) was also determined using STRUCTURE software and estimated as the average probability that a pair of alleles, one from each population A and B are distinct, less the average within-population heterozygosities [36].

A total of 12,886 SNPs with known physical positions were used for PCA in TASSEL 5 [5]and PCA plots were generated in statistical software R.3.3.3 (R development core team. 2011). PCA output data were visualized according to STRUCTURE results (using $K = 2$ and $K = 4$).

A kinship matrix was determined using Genomic Association and Prediction Integrated Tool (GAPIT) [26]using the same SNP data as for PCA analysis. Kinship (K, relationship coefficient) was determined using

VanRaden's method 1[45].

Analysis of Genetic Diversity and Phylogenetic Tree: To examine and compare the diversity among various asparagus germplasm lines, distinct parameters were used, such as Observed Heterozygosity (Ho), genetic diversity or Expected Heterozygosity (He), and Polymorphism Information Content (PIC). These parameters were calculated using Power Marker v 3.25 [27]. PIC was calculated to determine the probability of finding a polymorphism between a pair of random samples in the germplasm based on the following equation [4]:

$$\mathbf{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i^2 p_j^2$$

where p_i and p_j are the frequencies of i^{th} and j^{th} alleles for the selected marker, respectively.

Expected heterozygosity is the probability that two randomly selected alleles from a population are different and was calculated by the equation outlined by Harris & DeGiorgio, [18]:

$$\boldsymbol{He} = 1 - \sum_{i=1}^{n} p_i^2$$

where i represents the number of distinct alleles at a locus and p_i is the frequency of an allele in the population.

Analysis of Molecular Variance (AMOVA) within and among asparagus germplasm lines was conducted with the software GenAlEx v 6.5 [33]. The fixation index was further determined from AMOVA using GenAlEx v 6.502 [33].

Genetic distances (genetic divergence within a species or between populations within a species) of the 64 asparagus germplasm lines were calculated using Tassel v 5.2.43 [5] and estimated as 1 − IBS (Identity by State).

Phylogenetic trees were also constructed using genotypic data. Dendrograms were developed using a neighbour joining method in Tassel v 5.2.43. Plots were visualized with Dendroscope v 3.2.10 [19].

SNP Distribution and Linkage Disequilibrium Analysis: The density of SNPs over each chromosome was determined with the CM plot package in R [47]. Software TASSEL v 5.2.43 was used to calculate pair-wise LD, r^2 values between the SNP markers and the corresponding p-values for each r^2 estimate. Linkage disequilibrium decay was calculated in the population and visualized using the ggplot2 package in R.

3. Results

3.1. Summary description of SNP dataset

The 12,886 mapped SNPs covered 1,308 Mb of the asparagus genome, resulting in an average SNP density of 9.6 SNPs/Mb (Fig. 1). The number of SNPs on each chromosome ranged from 659 (Chromosome 9) to 1818 (Chromosome 1) (Table 1 and Additional File 1: Fig. S1). SNP density also varied among chromosomes; the highest was recorded on chromosome 4 (116 SNPs/kb) and lowest on chromosome 3 (67.8 SNPs/kb) (Table 1). The following SNP types were observed: [CT] (26.9 %), [GA] (26.5 %), [GT] (12.1 %), [CA] (12.5 %), [AT] (12.4 %), and [CG] (9.5 %). He, PIC value and Ho showed variation between chromosomes (Fig. 2). The trends of all three parameters on chromosomes were similar. PIC values ranged from 0.100 to 0.470 with a mean of 0.310. Approximately 58 % of SNPs had PIC estimates of > 0.3. The highest He, PIC, and Ho values were found on chromosome 2 and lowest were seen on chromosome 10 (Fig. 2). The Observed Heterozygosity, a measure of allelic diversity at a locus, ranged from 0.010 to 1 with a mean of 0.450. Approximately 58 % of the estimates were lower than 0.5, whereas He varied from 0.110 to 0.560 with an average of 0.370. Only 40 % of He estimates were greater than 0.450.

4. Population structure and kinship

The genetic distance based on pairwise comparison of 64 lines ranged from 0.266 to 0.365 with an overall average of 0.334, which suggests high genetic variation in the germplasm. The majority (89.3 %) of the genetic distances were greater than 0.320. The lowest genetic distance (0.266) was observed between cultivars NJ1122 (New Jersey) and NJ1113 (New Jersey). The genetic distance between PI 277830 (England) and Ravel (Germany) was the greatest at 0.366 (data not shown).

The delta K plot proposed that the germplasm panel could be separated into two, four, seven, and nine groups (Fig. 3a). Structure analysis demonstrated that Canadian lines formed a separate group for all four K

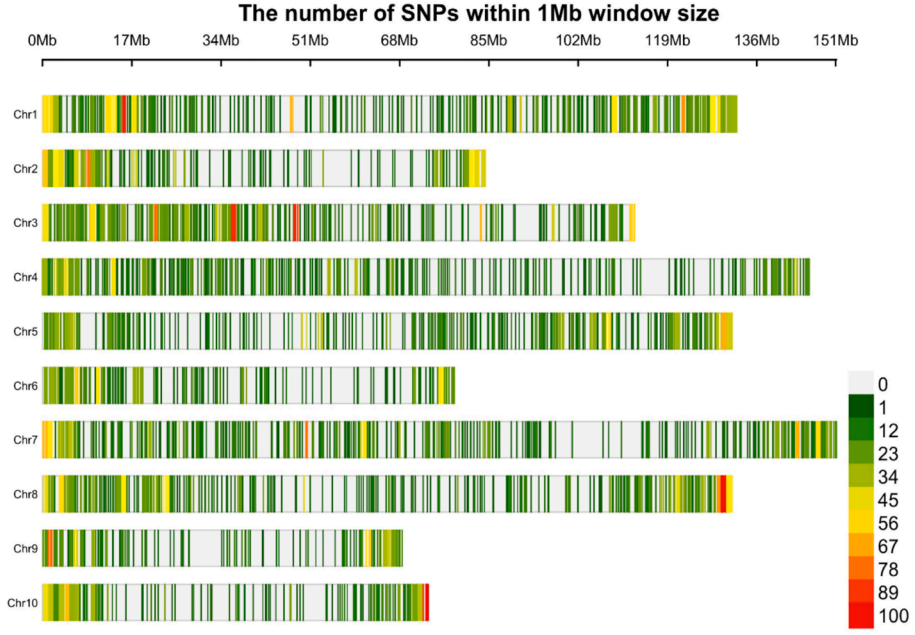

Fig. 1. Single Nucleotide Polymorphism density plot within 1 Mb window size over 10 asparagus chromosomes for 64 lines. The different colours represent SNP density, while the horizontal axis depicts the chromosome length (Mb).

Table 1
Summary for 12,886 SNP markers identified in 64 germplasm lines for the 10 asparagus chromosomes.

Chromosome	Number of SNPs	%SNPs on each chromosome	Start position on reference genome	End position on reference genome	Length (Size of chromosome in Mb)	Density (SNPs/ Mb)
1	1818	14.1	123,191	132,207,382	132.4	13.7
2	968	7.5	136,443	84,309,732	84.7	11.4
3	1670	13.0	91,303	112,682,426	113.3	14.7
4	1264	9.8	2684	146,001,793	146.6	8.6
5	1246	9.7	347,879	131,323,107	131.5	9.5
6	726	5.6	31,870	78,436,885	78.8	9.2
7	1670	13.0	59,978	151,152,493	151.4	11.0
8	1574	12.2	188,313	131,267,410	131.3	12.0
9	659	5.1	53,742	68,494,281	68.5	9.6
10	851	6.6	201,365	73,427,822	73.5	11.6

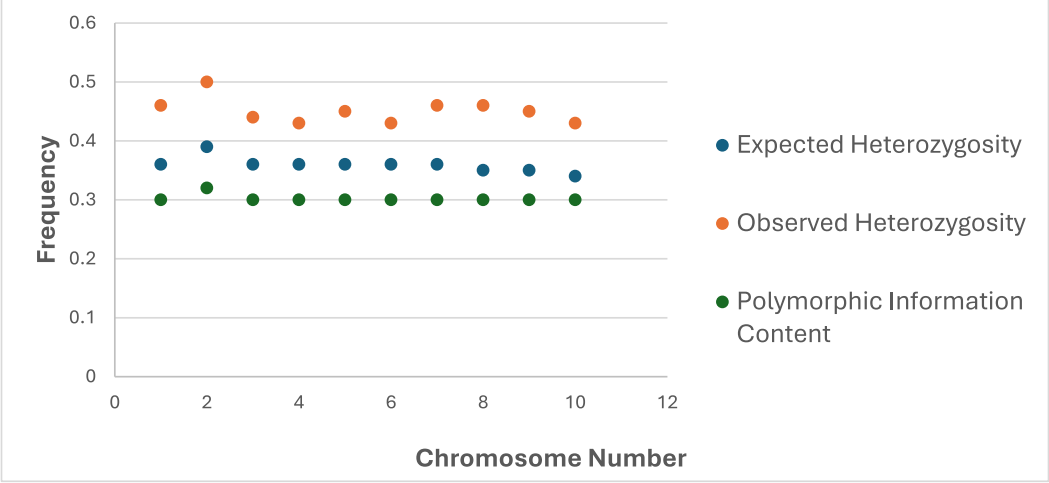

Fig. 2. Summary statistics of 12,886 SNPs used for genotyping 64 asparagus germplasm lines.

values (Fig. 3b). The highest Delta K score was for K = 9, followed by K = 7, dividing the germplasm into nine and seven subgroups, respectively. However, neither grouping had dominant populations, where the proportion of membership coefficient was above 0.50. Consequently, further analyses were conducted for the K = 4 grouping, since all 64 germplasm lines were grouped into one of the four subpopulations utilizing 0.50 as the likelihood value (Fig. 3b). For K = 4, the groupings included lines originating from: (1) New Zealand, New Jersey, France, and California; (2) Canada; (3) China, The Netherlands, and Germany; and (4) England, Denmark, Spain, Turkey, and India.

The PCA analysis, performed based on regions, lines from New Jersey, The Netherlands, and Canada formed separate groups (Fig. 4a). A PCA plot indicated clear separation of the Canadian lines from the remainder of the germplasm panel when K = 2 (Fig. 4b). Group 1 showed more dispersion than Group 2, which contained Canadian lines, indicating that Group 1 can be further divided into subgroups. For K = 4, Group 1, included lines from New Jersey, France, New Zealand, and California, Group 2 contained Canadian lines, Group 3 had lines from The Netherlands, China and Germany, whereas, Group 4 had lines from India, England, Denmark, Spain, Turkey (Fig. 4c). Group 3 and Group 4 were distinct, but with some admixtures.

Expected Heterozygosity within subpopulations ranged from 0.084 (Subpopulation 2- Canada) to 0.179 (Subpopulation 4- Denmark, Spain, England, India, Turkey) with an average of 0.130 (Table 2). The low value in Subpopulation 2 (Canada) suggested low genetic variation, which was expected since the lines were hybrids with a common parent. Subpopulation 2 showed both the highest F_{st} and lowest He, further indicating that lines in this subpopulation were more closely related to each other than to lines in other subpopulations. The high genetic variation in Subpopulation 4 can be explained by the presence of open-pollinated lines.

Subpopulation 4 had the lowest F_{st} value of 0.084 and the highest F_{st} was noted in sub-population 2 (0.517). Pairwise genetic distance in all subpopulations formed on basis of location varied from 0.001 (England with Spain) to 0.089 (California with Canada), suggesting a considerable degree of differentiation between the populations (Table 3).

The highest allele frequency divergence (0.047) was observed between subpopulations 3 (The Netherlands, China, and Germany) and 2 (Canada), whereas the lowest allele frequency divergence (0.021) was recorded between subpopulation 1 (New Jersey, France, New Zealand, and California) and subpopulation 4 (India, England, Denmark, Spain, and Turkey) (Table 4). Overall, these values between populations tended to be low, suggesting inter-breeding and low isolation.

The NJ phylogenetic tree displayed division of groups similar to those from population structure results based on country of origin (Fig. 5a). Lines from New Jersey, New Zealand, France, and California grouped together, as did those from China and The Netherlands, England and India, and Canada. When the lines were classified based on cultivation practice (green and white), clear separations were observed, but some displayed cross-over within the tree (Fig. 5b). Overall, results based on STRUCTURE analysis, PCA and the NJ phylogenetic tree were consistent.

For the kinship matrix, approximately 70 % of the coefficient estimates were equal to zero, suggesting low relatedness and little similarity for allele composition among most lines (Table 5). The highest kinship values were found between two lines from The Netherland which originated from same breeding program. The genetic relatedness (sharing alleles due to recent common ancestry) as analyzed by using Efficient Mixed Model Association (EMMA) software proved to be weak (Fig. 6), since most values were distributed between 0 and 0.5.

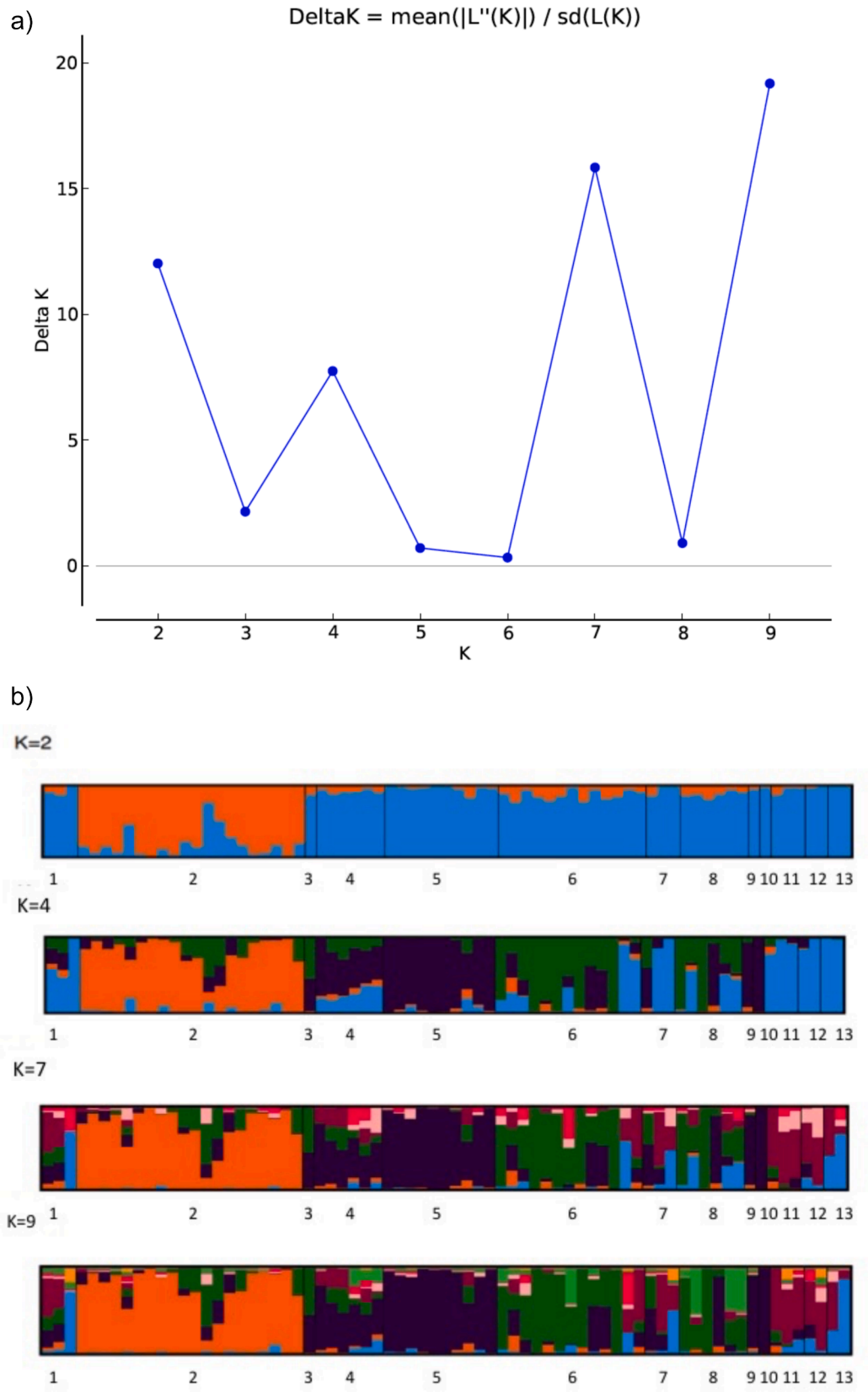

Fig. 3. Population structure results: a) Delta K plot of Evanno's test based on STRUCTURE analysis indicating population structure of 64 asparagus germplasm lines (K = number of subpopulations in the population, L" = the log likelihood of the data changes between successive K values); b) STRUCTURE plots indicating population structure of 64 asparagus germplasm lines derived from 12,886 SNPs. 1: Denmark, 2: Canada, 3: China, 4: New Zealand, 5: New Jersey, 6: The Netherlands, 7: Spain, 8: Germany, 9: France, 10: California, 11: England, 12: India, 13: Turkey.

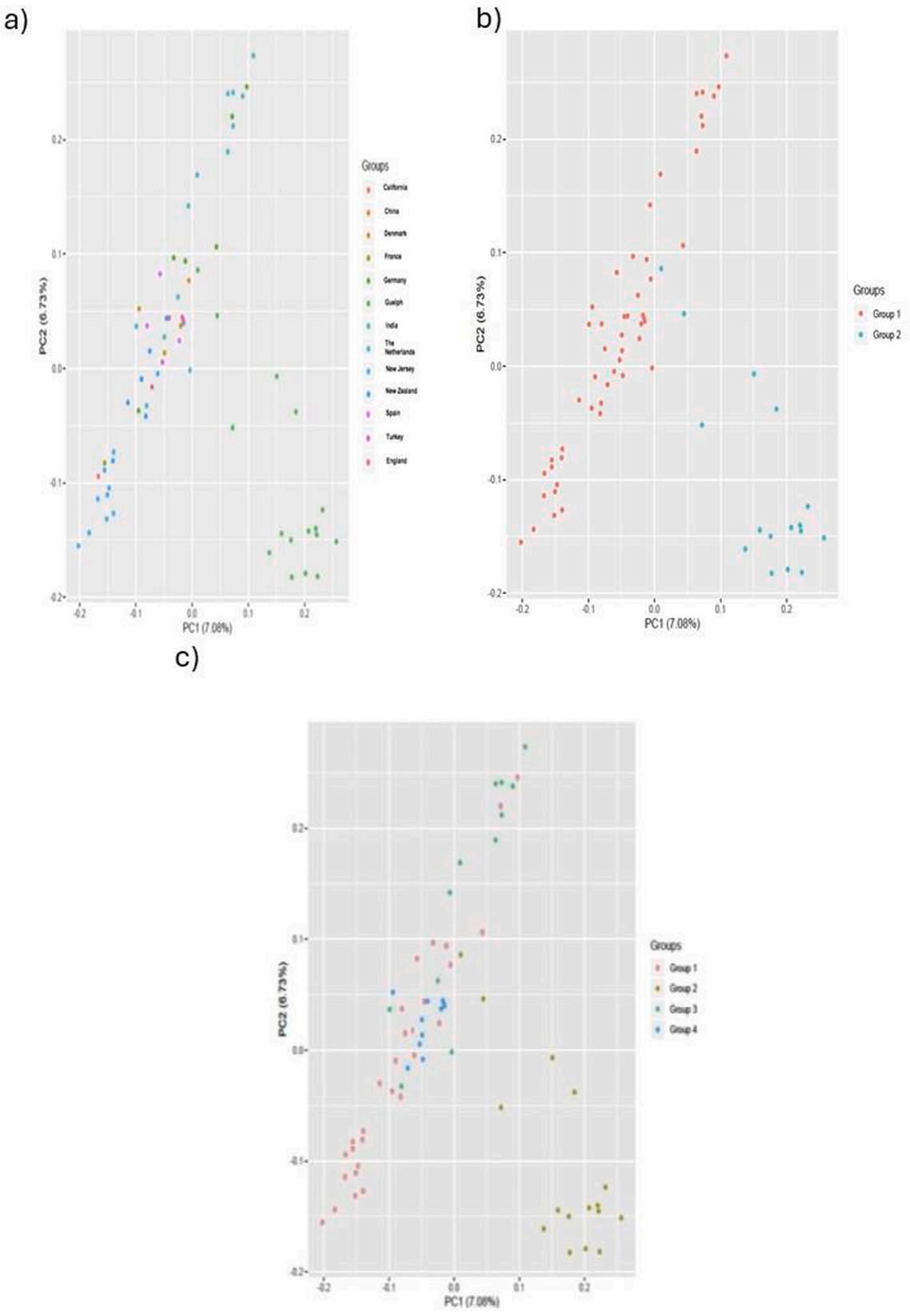

Fig. 4. PCA plots of 64 asparagus germplasm lines, colour coded according to: a) Country/Region; b) results of STRUCTURE software K = 2 [(1)-All lines excluding Canada, (2)-Canada]; c) K = 4 [(1) New Jersey, France, California, New Zealand (2) Canada (3) The Netherlands, Germany, China (4) England, India, Spain, Turkey, Denmark].

Table 2
Mean fixation index (F_{st}) and Expected Heterozygosity (He) for four subpopulations derived from 64 asparagus germplasm lines using STRUCTURE analysis, K = 4. The number of lines in each country are indicated in parentheses.

Subpopulations	Countries	F_{st}	He
1	California (3), New Jersey (9), France (1), New Zealand (6)	0.284	0.130
2	Canada (15)	0.517	0.086
3	The Netherlands (12), China (1), Germany (7)	0.340	0.125
4	India (1), England (2), Denmark (3), Spain (2), Turkey (2)	0.084	0.180

4.1. Genetic diversity

Expected Heterozygosity of all lines based on country of origin ranged from 0.110 to 0.560 with an average of 0.370 (data not shown). Among groups, Denmark showed the highest diversity (0.426) and France showed lowest diversity (0.220) (Table 6). Observed Heterozygosity in each group ranged from 0.287 (Turkey) to 0.528 (China). PIC values ranged from 0.165 (France) to 0.354 (Denmark).

Analysis of Molecular Variance results indicated that 4 % of total variation occurred among populations grouped based on STRUCTURE subpopulations, whereas the remaining 96 % of total variation was within populations (Table 7).

4.2. Linkage disequilibrium analysis

Linkage disequilibrium declined rapidly with an increase in physical distance (Fig. 7). The genome-wide LD decay distance was approximately 0.22 Mbp with an r^2 cut-off threshold of 0.2.

5. Discussion

Analysis of 12,886 GBS-derived SNP markers from 64 asparagus germplasm lines indicated high levels of genetic diversity associated with geographical region of line origin, and population mixing. Based on STRUCTURE analysis the lines were best divided into four groups: (1) New Zealand, New Jersey, France, and California; (2) Canada; (3) China, The Netherlands, and Germany; and (4) England, Denmark, Spain, Turkey, and India. Phylogenetic trees and PCA further supported these findings. Groups one and two had the highest, while groups one and three had the lowest, allele frequency divergence. Data are useful for predicting heterotic groupings in hybrid development.

To examine and compare diversity among genotypes, distinct parameters have been used such as Observed Heterozygosity, Expected Heterozygosity, and PIC. Markers with a PIC value of 0.5 are believed to be highly informative and best in differentiating genetic lines, whereas those with values 0.25 to 0.5 are moderately informative, and those with values less than 0.25 are marginally informative. Most markers in this study had PIC values greater than 0.25 (mean = 0.31), indicating they were sufficiently informative and could be used in genetic characterization studies to detect polymorphisms among asparagus genotypes. Markers with high PIC values can be specifically utilized to differentiate closely related genotypes. Results of this study demonstrated that many SNP markers were polymorphic and capable of differentiating asparagus genotypes despite the predicted narrow genetic base, in which most of the cultivars originated from single source[23].

Previously, genetic diversity was assessed by using six EST-SSR markers in cultivated diploid asparagus and a hexaploid population (Garcia et al., 2021a). The mean PIC value of the hexaploid population was higher than that for the diploid cultivars. In addition, the total number of alleles detected was higher for the hexaploid (81 alleles) than the diploids (25 alleles) which may have been predicted due to the undomesticated nature of the hexaploids and presence of duplicate loci. The mean PIC value of diploid cultivars, 0.615, reported as by Garcia et al. (2021a), was higher than that reported in this study, possibly as a result of the marker types studied, EST-SSR versus SNP markers, which are usually multiallelic and biallelic markers, respectively. The mean PIC value observed here for asparagus was lower than that determined analyzing SNPs in maize[48,21] and wheat but comparable to those in tomato [10], and soybean[8].

Expected Heterozygosity is the probability that two randomly selected alleles from a population are different[28]based on the number of alleles and their frequencies in the population with values ranging from 0 for no heterozygosity to 1 for maximum heterozygosity (high number of alleles with equal frequencies). The He of all lines based on country of origin varied from 0.110 to 0.560 with an average of 0.370. About 40 % of He estimates were greater than 0.450. Based on these values, the germplasm panel seems to be genetically diverse and can be used to develop new asparagus cultivars with desirable traits.

Observed Heterozygosity is another important measure of variability in a population. High Observed Heterozygosity suggests large genetic variability whereas low heterozygosity suggests little genetic variability is observed in the population. Observed Heterozygosity also ranges from 0 (no observed heterozygotes) to 1 (all heterozygotes). In this study Ho of all lines based on the country of origin varied from 0.010 to 1 with a mean of 0.450 suggesting outbreeding. Lower Expected Heterozygosity than Observed Heterozygosity also indicates the existence of an isolation-breaking effect (intermixing of divergent populations) [42]. Studies in other crops reported Observed Heterozygosity lower than Expected Heterozygosity[16,17], which normally suggests inbreeding

Table 4
Allele frequency divergence (net nucleotide distance) among four subpopulations derived from 64 asparagus germplasm lines using STRUCTURE analysis, K = 4. The number of lines in each country are indicated in parentheses.

Countries	Sub-populations	2	3	4
California (3), New Jersey (9), France (1), New Zealand (6)	1	0.045	0.037	0.021
Canada (15)	2	–	0.047	0.038
The Netherlands (12), China (1), Germany (7)	3		–	0.026
India (1), England (2), Denmark (3), Spain (2), Turkey (2)	4			–

Table 3
Pairwise genetic distance values between asparagus germplasm lines grouped by country of origin excluding subpopulations with less than two lines.

Subpopulation	Denmark	Canada	New Zealand	New Jersey	Germany	The Netherlands	California	Spain
Canada	0.067							
New Zealand	0.020	0.053						
New Jersey	0.047	0.063	0.024					
Germany	0.028	0.054	0.034	0.051				
The Netherlands	0.031	0.048	0.030	0.035	0.016			
California	0.066	0.089	0.033	0.036	0.070	0.053		
Spain	0.014	0.083	0.036	0.058	0.050	0.034	0.066	
England	0.016	0.077	0.041	0.062	0.042	0.040	0.074	0.001

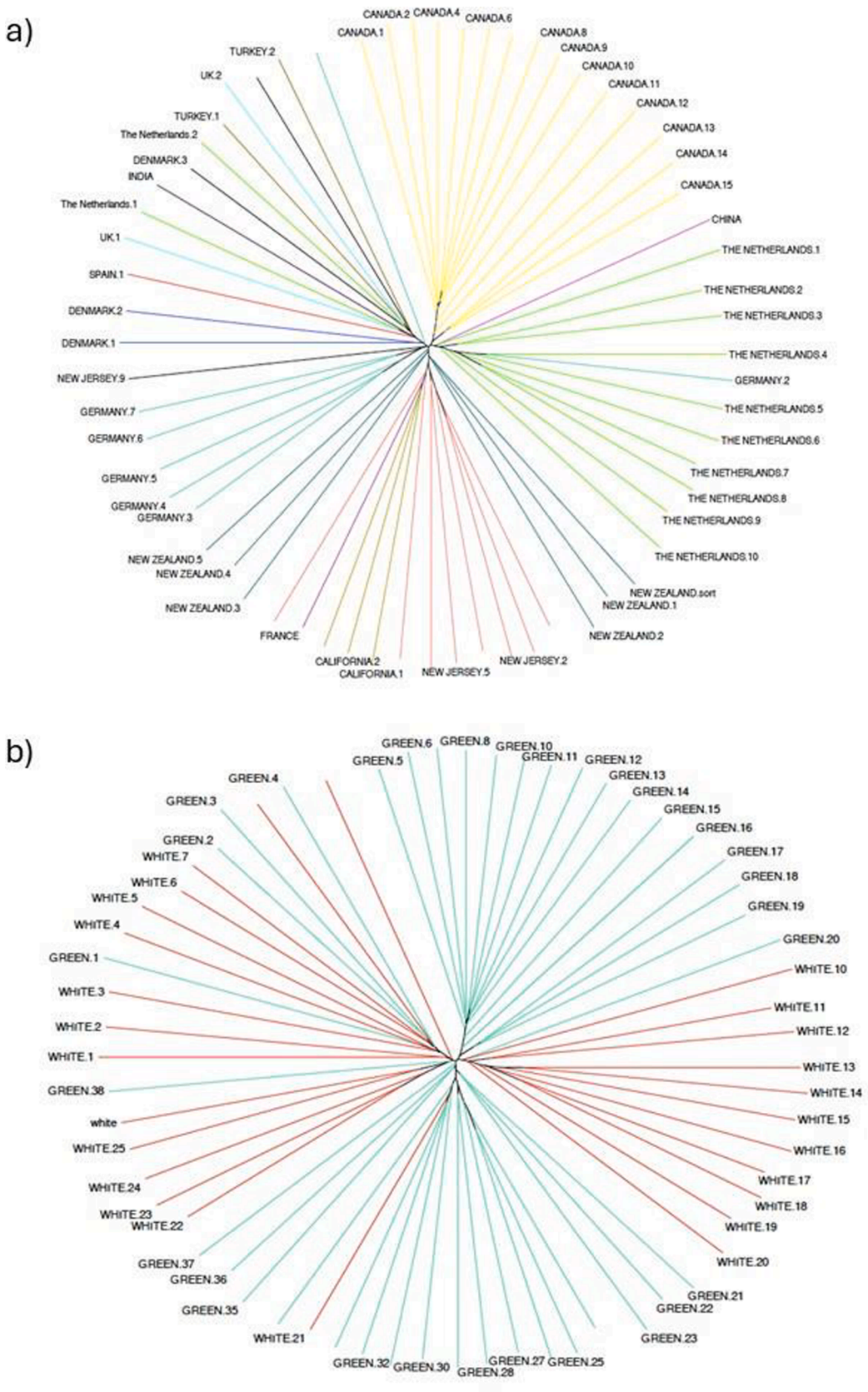

Fig. 5. The neighbor-joining phylogenetic tree of 64 asparagus germplasm lines based on a) breeding origin and b) cultivation practice- green or white. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Distribution of pairwise kinship between 64 asparagus germplasm lines.

Relative Kinship	Percentage
≤0.000	71.300
0.000–0.200	26.318
0.200–0.300	0.732
0.300–0.400	0.098
0.400–0.500	0.000
>0.500	1.563

and the presence of population structure among the germplasm[37].

Expected Heterozygosity provides an understanding of variability between individuals in the same subpopulation. Among subpopulations, Denmark showed the highest value (0.426) and France showed lowest diversity (0.220). Low levels were found in the single lines obtained from France, China (0.264), and India (0.223), where five plant were sampled for each and was likely an insufficient number to be representative of those heterogeneous populations (Table 3.6) One line in the subpopulation is not sufficient to recover within population genetic diversity estimates and values may have limited utility. Despite the fact that the single line from the France subpopulation is Argenteuil, an ancestral population, which was expected to have high diversity, increased heterozygosity levels were found in Denmark (0.426), because of increased population size that included three open-pollinated populations.

Four distinct sub-populations were defined for the 64 lines based on STRUCTURE analysis. The results were further confirmed by PCA and phylogenetic trees (Figs. 5 and 6). The majority of asparagus germplasm lines from The Netherlands and Germany clustered into a subgroup, which is expected, as lines from these countries are believed to be descendants from Braunschweig [23]. New Jersey, France and California lines were also clustered together. The line from France was Argenteuil and those from New Jersey and California are believed to derive from Mary Washington, which originated from Reading Giant (Argenteuil gene pool) [12,23,31]. These findings support most cultivars being

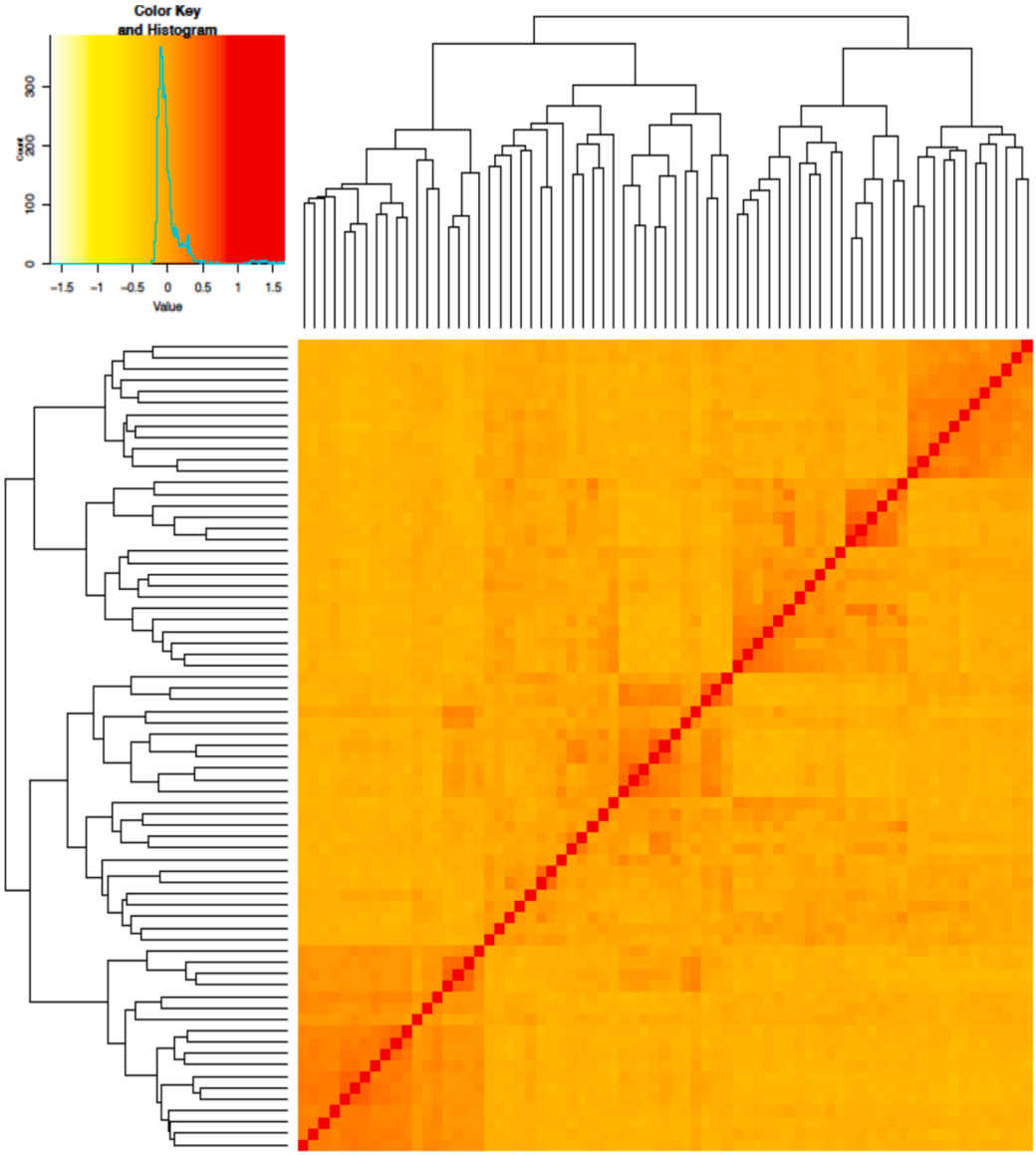

Fig. 6. Heatmap of kinship matrix values of 64 asparagus germplasm lines based on SNP markers from GAPIT according to the VanRaden algorithm. The Colour Key and Histogram displays the distribution of coefficient of coancestry. Strong red colour indicates lines that are most related. The dendrogram on the left and top displays the results of a cluster analysis on the IBD (Identity by Descent) matrix. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Expected Heterozygosity (He), Observed Heterozygosity (Ho) and Polymorphic Information Content (PIC) of different groups of asparagus germplasm lines based on region.

Group	Number of lines in each group	Expected Heterozygosity (He)	Observed Heterozygosity (Ho)	PIC value
New Jersey	9	0.318	0.449	0.256
France	1	0.220	0.440	0.165
California	3	0.233	0.466	0.175
China	1	0.264	0.528	0.198
Denmark	3	0.426	0.386	0.354
Germany	7	0.328	0.453	0.263
Canada	15	0.319	0.468	0.257
India	1	0.223	0.447	0.168
The Netherlands	12	0.341	0.439	0.278
New Zealand	6	0.330	0.475	0.264
Spain	2	0.325	0.429	0.259
Turkey	2	0.352	0.287	0.281
England	2	0.320	0.499	0.252

Table 7
Summary AMOVA of 64 asparagus germplasm lines from GenALEx.

Source	df	SS	MS	Estimated Variance	%
Among pops	3	3933.63	1311.2	34.0	4 %
Within pops	63	52533.4	784.1	784.1	96 %
Total	66	56,467		818.1	100 %
F_{st}	0.150				

df: Degrees of Freedom, SS: Sum of Squares, MS: Mean Square.

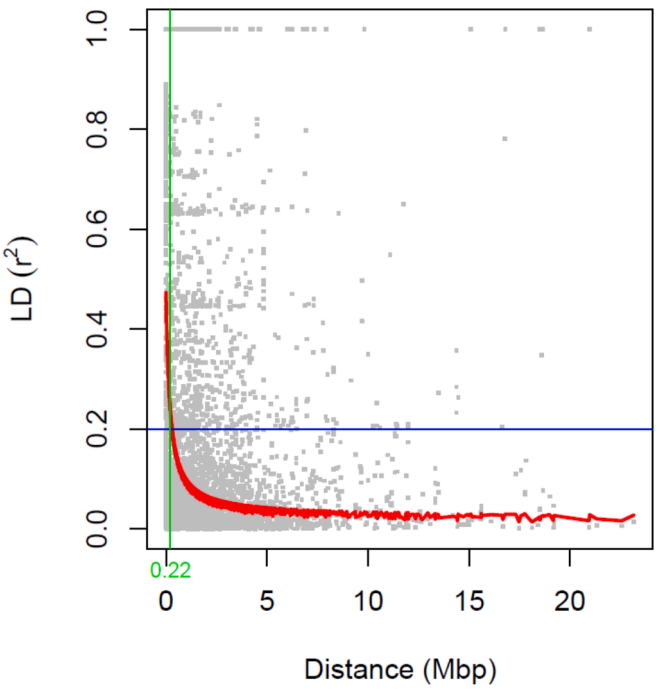

Fig. 7. Genome-wide linkage disequilibrium decay (LD) plot for 760 asparagus genotypes using 12,886 SNP markers. LD, measured as r^2, is plotted against the genetic distance (Mbp) (Threshold set at 0.2 (blue line)), the red curve line demonstrates the LD decay based on smoothing spline regression and the vertical green line indicates the LD decay (Mbp) at $r^2 = 0.2$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

derived from either of two populations: the Argenteuil population giving rise to cultivars from United States, France and Italy and the Braunschweiger population giving rise to cultivars from Germany and The Netherlands. The divergence is also supported by a study with 26 asparagus cultivars of eight different origins (France, Spain, Italy, USA, Taiwan, Germany, The Netherlands, and Turkey) by Geoffriau et al. [15] using morphological and isoenzyme markers. Agromorphological data divided the genotypes into two different clusters, one belonging to France, Germany, and The Netherlands, and the other belonging to the United States. Yet in another study, genetic diversity was analyzed in 30 asparagus cultivars using EST-SSR markers [7] where both trees and PCA analyses divided the genotypes into two clusters based on allele frequencies: one comprised of cultivars from US, Spain, Italy and Germany, whereas the other included cultivars from the Netherlands, Germany, Spain and New Zealand [7]. This further indicated that asparagus breeders from US and Europe have used different germplasm groups with some cultivars having complex origin with alleles from both groups. Other diversity studies in asparagus have provided conflicting conclusions in which they failed to divide the germplasm based on their country of origin[9,1,2,25,32]. Poor clustering may be related to the choice of germplasm studied, which might be open pollinated in nature and very genetically diverse, thus preventing a clear differentiation of germplasm based on origin.

In this study, admixture of lines was also observed, likely resulting from breeding programs intercrossing distinct germplasm sources. Alternatively, grouping lines from different locations could result from selection for similar phenotypes. For example, in subpopulation 1, New Jersey and California lines, which are majorly grown for green colour and prefer warm climates, were grouped together. All Canadian lines were grouped together likely because they were hybrids sharing a common female parent and supermales from a limited number of sources.

The fixation index is a measure of population differentiation that explains genetic diversity because of population sub-division. In terms of population differentiation, a value larger than 0.150 is deemed significant. Therefore, significant divergence was observed among germplasm lines in this study (Table 3.2). However, in one subpopulation, a low value (Subpopulation 4: 0.084) was observed suggesting low genetic differentiation among the germplasm lines from India, England, Denmark, Spain, and Turkey. The values can also be affected by the sample size. Thus, some of these germplasms had low sample size that could have affected estimates. Low values (0.150) between the four subpopulations also agree with the AMOVA results, in which most of the variation (96 %) was observed within subpopulations and only 4 % was determined among subpopulations (Table 3.7).

Pairwise genetic distance among all populations based on country of origin varied from 0.001 (England with Spain) to 0.089 (California with Canada) (Table 3.3). The low differentiation between England and Spain suggests a high level of breeding between them. This was also supported by STRUCTURE results with England and Spain grouping in subpopulation 4. Maximum pairwise genetic distance was found in California and Canada suggesting considerable degree of differentiation between the populations. Similar results were also reported by STRUCTURE, with Canada in subpopulation 2 and California in subpopulation 1.

Allele frequency divergence (net nucleotide distance) is the divergence among populations. It is the average pairwise variation between alleles from different population in addition to the variation within each population [36]. Distance is estimated as 0 for similar populations. The highest allele frequency divergence (0.047) was recorded between subpopulations 2 (Canada) and 3 (The Netherlands, China and Germany), whereas the lowest allele frequency divergence (0.021) was recorded between subpopulations 1 (New Zealand, New Jersey, France, California) and 4 (Denmark, Spain, India, Turkey, England) suggesting that subpopulations 1 and 4 are most similar to each other, probably because they are derivatives from the same germplasm source (Table 3.4). Comparable results were seen in the population STRUCTURE analysis in which lines from Canada and India formed separate groups and those from Turkey and India were clustered in one group. Investigations of genetic diversity and evaluation of germplasm resources are required to understand the degree of phylogenetic relatedness.

To develop suitable hybrid combinations, lines from unrelated groups should be crossed as they could have improved combining ability as compared to crosses made among lines within the related groups. Various genetic diversity parameters, such as fixation index can help to determine heterotic groups. In the future, hybrids could be developed by crossing: (1) Canadian lines with those from California, Spain, New Jersey and England; (2) California lines with those from Denmark, England, Spain and Germany; and (3) English lines with those from New Jersey.

One major limitation in this study was the availability of only the first draft of the available reference genome with no information of how much it is related to the germplasm used in this study. Germplasm lines could be related to each other but their relationship to the reference genome may have been minimal.

For effective and successful breeding, well-characterized genetic material is required. Few genetic diversity studies have been conducted in asparagus based on molecular markers thus results reported here can serve as a basis for future research. A considerable amount of genetic diversity was observed in this study, despite previous suggestions that asparagus cultivars may have originated from one open-pollinated population. For future research, increasing the number of germplasm lines in each geographical area will improve genetic diversity estimates and provide important information for breeders.

The substantial genetic diversity observed in this study has major implications for asparagus breeding in the future, including the improvement of hybrid vigor, disease resistance, and overall cultivar performance. The discovery of discrete subpopulations provides a valuable tool for developing efficient breeding strategies. The identification of molecular markers in these subpopulations can further enhance selection effectiveness. With this knowledge, asparagus breeders can create superior cultivars that satisfy the needs of a shifting agricultural environment, guaranteeing continued success of the global asparagus industry.

6. Conclusion

Genome-wide SNPs were identified using a high throughput GBS procedure for 64 asparagus germplasm lines and used to determine genetic diversity, population structure and LD pattern. Population structure revealed that all lines could be grouped into two, four, seven, or nine sub-populations. At $K = 7$ and $K = 9$, there were groups that do not appear to have any dominant populations, where the proportion of membership coefficient was above 0.5. Limiting the analysis to $K = 4$ provided the most meaningful distribution for analysis because all four groups had some germplasm lines representing them. For $K = 4$, four distinct groups were defined: (1) New Zealand, New Jersey, France, and California; (2) Canada; (3) China, The Netherlands, and Germany; and (4) England, Denmark, Spain, Turkey, and India. These results were further confirmed by PCA and phylogenetic analysis. The diversity parameters of the germplasm panel revealed groupings that could result in heterosis when crossed. In addition, understanding of population structure and genetic variation in asparagus germplasm will assist in association mapping, MAS and genomic selection studies in the future.

7. Declarations

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Funding: This work was supported by Fox Seeds, Asparagus Farmers of Ontario, Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development Grants, and Ontario Ministry of Agriculture, Food, and Rural Affairs.

Authors' contributions: GS: Conceptualization, Investigation, Methodology, Data Analysis, Writing- Original draft and review and editing. DW: Conceptualization, Funding acquisition, Project administration, Supervision, Writing- review and editing.

CRediT authorship contribution statement

Gurleen Sidhu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Travis Banks:** Writing – review & editing, Visualization, Software, Methodology, Conceptualization. **David Wolyn:** Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dr. Gregory Perry for their suggestions and assistance in data analysis.

Authors' information: Department of Plant Agriculture, University of Guelph, Guelph, Canada

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgeb.2025.100491.

References

[1]. Ahmad N, Tian R, Lu J, et al. DNA fingerprinting and genetic diversity analysis in *Asparagus officinalis* L. cultivars using microsatellite molecular markers. *Genet Resour Crop Evol*. 2023;70(4):1163–1177.
[2]. Amato LD, Martin EA, López-Anido FS. Combining ability and heterosis for market yield in green asparagus. *Horticulturae*. 2022;8(6):489.
[3]. Amato LD, López-Anido FS, Zayas A, Martin EA. Genetic resources in asparagus: diversity and relationships in a collection from different origins and breeding status. *N Z J Crop Hortic Sci*. 2023;51(1):69–80.
[4]. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 1980;32(3):314–331.
[5]. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*. 2007;23(19):2633–2635. https://doi.org/10.1093/bioinformatics/btm308.
[6]. Burrell AM, Pepper AE, Hodnett G, et al. Exploring origins, invasion history and genetic diversity of *Imperata cylindrica* (L.) P. Beauv. (Cogongrass) in the United States using genotyping by sequencing. *Mol Ecol*. 2015;24(9):2177–2193. https://doi.org/10.1111/mec.13167.
[7]. Caruso M, Federici CT, Roose ML. EST–SSR markers for asparagus genetic diversity evaluation and cultivar identification. *Mol Breed*. 2008;21:195–204.
[8]. Chander S, Garcia-Oliveira AL, Gedil M, et al. Genetic diversity and population structure of soybean lines adapted to sub-Saharan Africa using single nucleotide polymorphism (SNP) markers. *Agronomy*. 2021;11(3):604.

[9]. Chen H, Guo A, Wang J, et al. Evaluation of genetic diversity within asparagus germplasm based on morphological traits and ISSR markers. *Physiol Mol Biol Plants*. 2020;26(2):305–315. https://doi.org/10.1007/s12298-019-00738-5.
[10]. Corrado G, Piffanelli P, Caramante M, Coppola M, Rao R. SNP genotyping reveals genetic diversity between cultivated landraces and contemporary varieties of tomato. *BMC Genomics*. 2013;14:1–14.
[11]. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour*. 2012;4(2):359–361. https://doi.org/10.1007/s12686-011-9548-7.
[12]. Ellison JH. Asparagus breeding. Breeding vegetable crops. Westport: AVI; 1986.
[13]. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol Ecol*. 2005;14(8):2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x.
[14]. Garcia V, Castro P, Turbet-Delof M, Gil J, Moreno R. Development and diversity analysis of an hexaploid pre-breeding asparagus population with introgressions from wild relative species. *Sci Hortic*. 2021;287, 110273.
[15]. Geoffriau E, Denoue D, Rameau C. Assessment of genetic variation among asparagus (*Asparagus officinalis* L.) populations and cultivars: Agromorphological and isozymic data. *Euphytica*. 1991;61(3):169–179. https://doi.org/10.1007/BF00039655.
[16]. Gopaulchan D, Motilal LA, Bekele FL, et al. Morphological and genetic diversity of cacao (*Theobroma cacao* L.) in Uganda. *Physiol Mol Biol Plants*. 2019;25:361–375.
[17]. Gwali S, Vaillant A, Nakabonge G, et al. Genetic diversity in shea tree (Vitellaria paradoxa subspecies nilotica) ethno-varieties in Uganda assessed with microsatellite markers. *Forests Trees Livelihoods*. 2015;24(3):163–175.
[18]. Harris AM, DeGiorgio M. Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. *Hum Biol*. 2017;89(1):21–46. https://doi.org/10.13110/humanbiology.89.1.02.
[19]. Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst Biol*. 2012;61(6):1061–1067. https://doi.org/10.1093/sysbio/sys062.
[20]. Jiang C, Lewis ME, Sink KC. Combined RAPD and RFLP molecular linkage map of asparagus. *Genome*. 1997;40(1):69–76. https://doi.org/10.1139/g97-009.
[21]. Josia C, Mashingaidze K, Amelework AB, Kondwakwenda A, Musvosvi C, Sibiya J. SNP-based assessment of genetic purity and diversity in maize hybrid breeding. *PLoS One*. 2021;16(8), e0249505.
[22]. Kapoor M, Mawal P, Sharma V, Gupta RC. Analysis of genetic diversity and population structure in Asparagus species using SSR markers. *J Genet Eng Biotechnol*. 2020;18(1):50.
[23]. Knaflewski M. Genealogy of asparagus cultivars. *Acta Hortic*. 1996;415:87–92. https://doi.org/10.17660/ActaHortic.1996.415.13.
[24]. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
[25]. Li S, Zhang G, Li X, et al. Genome-wide identification and validation of simple sequence repeats (SSRs) from Asparagus officinalis. *Mol Cell Probes*. 2016;30(3):153–160.
[26]. Lipka AE, Tian F, Wang Q, et al. GAPIT: Genome association and prediction integrated tool. *Bioinformatics*. 2012;28(18):2397–2399. https://doi.org/10.1093/bioinformatics/bts444.
[27]. Liu K, Muse SV. PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics*. 2005;21(9):2128–2129. https://doi.org/10.1093/bioinformatics/bti282.
[28]. Luo Z, Brock J, Dyer JM, et al. Genetic diversity and population structure of a Camelina sativa spring panel. *Front Plant Sci*. 2019;10. https://doi.org/10.3389/fpls.2019.00184.
[29]. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. Article 1 *EMBnet Journal*. 2011;17(1). https://doi.org/10.14806/ej.17.1.200.
[30]. Mascher M, Muehlbauer GJ, Rokhsar DS, et al. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J*. 2013;76(4):718–727. https://doi.org/10.1111/tpj.12319.
[31]. Mercati F, Riccardi P, Harkess A, et al. Single nucleotide polymorphism-based parentage analysis and population structure in garden asparagus, a worldwide genetic stock classification. *Mol Breed*. 2015;35(2):59. https://doi.org/10.1007/s11032-015-0217-5.
[32]. Mousavizadeh SJ, Gil J, Castro P, Hassandokht MR, Moreno R. Genetic diversity and phylogenetic analysis in Asian and European Asparagus subgenus species. *Genet Resour Crop Evol*. 2021;68:3115–3124.
[33]. Peakall R, Smouse PE. genalex 6: Genetic analysis in Excel. population genetic software for teaching and research. *Mol Ecol Notes*. 2006;6(1):288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x.
[34]. Pegiou E, Mumm R, Acharya P, de Vos RC, Hall RD. Green and white asparagus (*Asparagus officinalis*): a source of developmental, chemical and urinary intrigue. *Metabolites*. 2019;10(1):17.
[35]. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155(2):945–959. https://doi.org/10.1093/genetics/155.2.945.
[36]. Pritchard JK, Wen X, Falush D. Documentation for structure software: Version 2.3. Univ. Chicago 2002;10.
[37]. Rani R, Raza G, Tung MH, et al. Genetic diversity and population structure analysis in cultivated soybean (Glycine max [L.] Merr.) using SSR and EST-SSR markers. *PLoS One*. 2023;18(5), e0286099.
[38]. Riccardi P, Casali PE, Mercati F, Falavigna A, Sunseri F. Genetic characterization of asparagus doubled haploids collection and wild relatives. *Sci Hortic*. 2011;130(4):691–700.
[39]. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, WGS500 Consortium, Lunter G. Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 2014;46(8):912–918.
[40]. Sala T, Puglisi D, Ferrari L, et al. Genome-wide analysis of genetic diversity in a germplasm collection including wild relatives and interspecific clones of garden asparagus. *Front Plant Sci*. 2023;14, 1187663.
[41]. Sarabi B, Hassandokht MR, Hassani ME, Masoumi TR, Rich T. Evaluation of genetic diversity among some Iranian wild asparagus populations using morphological characteristics and RAPD markers. *Sci Hortic*. 2010;126(1):1–7. https://doi.org/10.1016/j.scienta.2010.05.028.
[42]. Sharma R, Kumar B, Arora R, Ahlawat S, Mishra AK, Tantia MS. Genetic diversity estimates point to immediate efforts for conserving the endangered Tibetan sheep of India. *Meta Gene*. 2016;8:14–20. https://doi.org/10.1016/j.mgene.2016.01.002.
[43]. Tomassoli L, Tiberini A, Vetten H-J. Chapter 10—Viruses of Asparagus. In G. Loebenstein & H. Lecoq (Eds.). Adv Virus Res (Vol. 84, pp. 345–365). Academic Press; 2012. Doi: 10.1016/B978-0-12-394314-9.00010-5.
[44]. Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinf*. 2017;18(1):5. https://doi.org/10.1186/s12859-016-1431-9.
[45]. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. 2008;91(11):4414–4423. https://doi.org/10.3168/jds.2007-0980.
[46]. Weeks NT, Luecke GR. Optimization of SAMtools sorting using OpenMP tasks. *Clust Comput*. 2017;20:1869–1880.
[47]. Yin Y, Jiang Y, Lam K-W-G, et al. High-throughput single-cell sequencing with linear amplification. *Mol Cell*. 2019;76(4):676–690.e10. https://doi.org/10.1016/j.molcel.2019.08.002.
[48]. Zhang X, Zhang H, Li L, et al. Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. *BMC Genomics*. 2016;17(1):697. https://doi.org/10.1186/s12864-016-3041-3.