Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Genome-wide characterization of vegetable soybean core collection reveals genetic diversity for breeding efforts in Sub-Saharan Africa

Mahoussi Kadoukpe Arnaud Djanta
Makerere University

Eric Etchikinto Agoyi
Université d'Abomey-Calavi

Emmanuel O. Omondi
World Vegetable Center

Tonny Obua
Makerere University

Thomas L. Odong
Makerere University

Isaac Onziga Dramadri
Makerere University

Mildred Ochwo-Ssemakula
Makerere University

Ephraim Nuwamanya
Makerere University

Paul Gibson
Makerere University

Florent Jean-Baptiste Quenum
Université d'Abomey-Calavi

Richard Edema
Makerere University

Achille Assogbadjo
Université d'Abomey-Calavi

Ramakrishnan M. Nair
World Vegetable Center South and Central Asia

Phinehas Tukamuhabwa
Makerere University

Ya-Ping Lin
ya-ping.lin@worldveg.org

World Vegetable Center

Research Article

Keywords: vegetable soybean, genebank core collection, population structure, genetic diversity, Sub-Saharan Africa

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10573993/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Vegetable soybean [*Glycine max* (L.) Merr] is an underutilized legume with considerable potential to improve nutrition and diversify cropping systems in Sub-Saharan Africa (SSA). To support its genetic improvement, a soybean panel including the World Vegetable Center large-seeded soybean core collection and SSA elite grain soybeans was characterized using genome-wide single-nucleotide polymorphism (SNP) markers. Genetic groups were investigated using principal component analysis (PCA), population structure, discriminant analysis of principal components (DAPC), and phylogenetic analysis. Genome-wide analyses consistently resolved the collection into three major genetic groups. Population differentiation reflected both market class and breeding history, with two clusters grouping predominantly vegetable soybean accessions and one cluster containing mainly SSA elite grain soybeans. Analysis of molecular variance indicated that 83.78% of the total genetic variation resided within populations, whereas 16.22% occurred among populations. Cluster 1 exhibited the highest genetic diversity (PIC = 0.253; He = 0.262), whereas cluster 3, comprising predominantly SSA elite grain soybeans, showed comparatively lower diversity (PIC = 0.188; He = 0.178), indicating a relatively narrow genetic base. Pairwise F_{ST} and Nei's genetic distance further demonstrated moderate genetic differentiation among the three populations. The results demonstrate that the World Vegetable Center large-seeded soybean core collection represents a valuable reservoir of genetic diversity for vegetable soybean improvement. The identified population structure and diversity patterns provide a useful foundation for germplasm conservation, parental selection, and the development of vegetable soybean cultivars adapted to Sub-Saharan Africa.

Introduction

Plant genetic resources comprise germplasm materials that serve as a reservoir of genetic diversity for breeding improved varieties with desired traits (Gueco et al. 2022). Among legume crops, soybean [*Glycine max* (L.) Merril (2n = 40)] is one of the world’s most economically important legume crops, providing a significant supply of plant protein and oil for food, feed, and industrial use (Nguyen and Bhattacharyya 2017; Shaibu et al. 2022). Beyond its traditional use as a mature grain crop, the soybean is also harvested at the immature pod stage and consumed as a vegetable (Huang et al. 2022; Nair et al. 2023). Vegetable soybeans represent around 2% of total soybean production in the world (Moinuddin et al. 2023). Edible or green or vegetable soybeans are prized for bigger seed size, sweetness, tenderness, and nutritional content (Agyenim-Boateng et al. 2023; Soujanya et al. 2025; Meghana et al. 2025). Although belonging to the same species, grain and vegetable soybeans represent different market segments influenced by diverse selection pressures regarding pod traits, seed size, as well as sensorial attributes (Dong et al. 2014; Moseley et al. 2021; Yu et al. 2022). Vegetable soybean breeding targets large seeds, attractive green pods, large beans, sweetness, tenderness, and superior eating quality to suit fresh vegetable markets (Moseley et al. 2021; Nair et al. 2022). Conversely, grain soybean breeding aims to enhance grain yield, seed protein, and oil content, as well as biotic and abiotic stress tolerance (Hartman et al. 2011; Fang et al. 2024).

Cultivated soybean was domesticated from its wild progenitor, *Glycine soja*, through strong artificial selection for key agronomic traits, including reduced pod shattering, increased seed size, and adaptation to diverse agricultural environments, resulting in substantial genomic changes during domestication (Hyten et al. 2006; Li et al. 2010; Sedivy et al. 2017). Consequently, soybean cultivars exhibit considerably lower genetic diversity than their wild ancestor because of domestication bottlenecks and subsequent breeding based on a limited number of elite founder parents (Hyten et al. 2006; Zhuang et al. 2022). Following domestication, soybeans expanded into diverse agroecological regions where adaptation to local climatic conditions and agricultural practices promoted genetic differentiation among geographically distinct populations (Wang et al. 2016; Valliyodan et al. 2021). Population structure analyses have consistently revealed geographically structured subpopulations that reflect the combined effects of historical dispersal, local adaptation, and breeding (He et al. 2016; Valliyodan et al. 2021). For example, selection on flowering- and maturity-related genes, including the major *E* loci and the *J* locus, facilitated adaptation to high-latitude and tropical environments and contributed to genetic differentiation among regional soybean populations (Jiang et al. 2014; Lu et al. 2017; Dong et al. 2021). Genomics studies further demonstrated that genetic differentiation is unevenly distributed across the soybean genome, with specific genomic regions exhibiting signatures of selection associated with domestication, regional adaptation, and improvement (Zhou et al. 2015; Maldonado Dos Santos et al. 2022). These selection processes have altered allele frequencies among soybean market classes, particularly between grain and vegetable soybeans, reflecting divergent breeding objectives and end-use preferences (Liu et al. 2022, 2025a). Characterizing the genetic diversity and population structure of soybean germplasm is fundamental for effective conservation, association mapping, and the identification of genetically complementary parents for breeding. Early studies based on low-density molecular markers such as single sequence repeats (SSRs) and amplified fragment length polymorphism (AFLP), provided valuable insights into soybean diversity and domestication history (Powell et al. 1996; Hyten et al. 2006). However, the limited marker density and uneven genome coverage constrained their ability to resolve fine-scale population structure and genetic relationships (Myles et al. 2009; Song et al. 2013). The advent of high-density genome-wide SNP genotyping has significantly improved the resolution of population genetic analyses by providing comprehensive coverage of allelic variation across the genome (Song et al. 2013; Zhou et al. 2015; Niu et al. 2024). Such datasets enable more accurate estimation of genetic diversity, population differentiation, admixture, and relatedness, thereby facilitating the efficient conservation and utilization of soybean genetic resources and informing parental selection for modern breeding programs (Silue et al. 2025; Song et al. 2013; Viana et al. 2024).

Soybean production in Sub-Saharan Africa (SSA) has significantly expanded over the past three decades, driven by the release of improved grain soybean varieties developed through breeding efforts led by the International Institute of Tropical Agriculture (IITA) and collaborating research institutions (Tukamuhabwa and Oloka 2016; Khojely et al. 2018; Chigeza et al. 2019). Continental multi-environment evaluation trials have further facilitated the identification of high-yielding and broadly adapted grain soybean germplasm, contributing to the development of breeding resources for African production systems (Agoyi et al. 2024; Araújo et al. 2025). In contrast, vegetable soybean has received relatively

less breeding attention despite its nutrient value and potential to diversify legume-based food systems and alleviate micronutrient deficiencies (Agyenim-Boateng et al. 2023; Soujanya et al. 2025; Meghana et al. 2025). The adoption of vegetable soybean in SSA remains limited because most introduced germplasm exhibits poor adaptation, low yield stability, and inadequate agronomic performance under African environments (Ahomondji et al. 2022; Tsindi et al. 2019; Van Der Merwe et al. 2024). Therefore, developing large-seeded vegetable soybeans with broad adaptation and stable performance is essential for establishing sustainable production and seed systems across the region.

To support this objective, the World Vegetable Center recently developed a representative core collection of large-seeded soybean germplasm (Nair et al. 2022). However, the genetic diversity, population structure, and evolutionary relationships within this collection have not been comprehensively characterized using genome-wide markers. Such knowledge is critical for identifying genetically diverse parental lines, maximizing allelic diversity in breeding populations, and facilitating the efficient utilization of germplasm for vegetable soybean improvement in Africa. Therefore, this study implemented high-density SNP genotyping to characterize the genetic diversity, population structure, and phylogenetic relationships of the World Vegetable Center large-seeded soybean core collection, with other elite breeding materials from local institutes, providing a genomic foundation for germplasm conservation and breeding of vegetable soybean adapted to Sub-Saharan Africa.

Materials and Methods

Plant materials

A collection of 191 soybean accessions from five research institutes and two seed companies was used in this study (Fig. 1 and Table S1). This collection included 109 vegetable soybean core accessions from the World Vegetable Center (Nair et al. 2022), 26 vegetable soybeans from Seed Savers Exchange, United States of America, 3 vegetable soybeans from the National Agricultural Research Organization, Japan, and 53 elite grain soybeans sourced directly from the Pan African Variety trials initiative (Agoyi et al. 2024; Araújo et al. 2025). These SSA grain soybeans comprised accessions from the International Institute of Tropical Agriculture (37), from the Council for Scientific and Industrial Research (5), from Makerere University (1), and the seed company Seed Co Limited (10).

Restriction site-associated DNA (RAD) sequencing

Plants were grown in pots under screenhouse conditions at the Faculty of Agricultural Sciences of the University of Abomey-Calavi, Benin (9°18’27.7”N, 2°18’57”E). At 15 days after sowing, three pieces of young leaves per accession were collected and placed in 15 mL tubes containing silica gel, then sealed and labeled. The samples were packed in zipper bags, stored in a carton box, and shipped to the Biotechnology Laboratory of the World Vegetable Center, Taiwan, for DNA extraction.

Genomic DNA was extracted using the FavorPrep Plant Genomic DNA Extraction Mini Kit (Favorgen Biotech Corp., Pingtung, Taiwan), following the manufacturer's protocol. The DNA library was prepared using the method developed by Elshire et al. (2011). In brief, DNA was quantified using Qubit and standardized to 100 ng per sample. "*Ape*KI" (New England Biolabs, Ipswich, MA, USA) was used to digest the DNA samples, which were subsequently ligated with two sequencing adapters and amplified by polymerase chain reaction to produce target DNA fragments. The sequencing library preparation was then completed by selecting fragments that fell within 300 bp to 500 bp. Libraries were sequenced using the Illumina platform on a NovaSeq 6000 system with 150 bp paired-end reads.

SNP development

After obtaining raw sequencing data, process_radtags were used in Stacks (version 2.68) to examine and demultiplex the reads (Catchen et al. 2013; Rochette et al. 2019). Reads containing ambiguous barcodes, uncalled bases (Ns), or missing or unrecognized restriction enzyme cut sites were discarded. Quality filtering was performed using the default parameters implemented in process_radtags, which remove reads failing the built-in quality control criteria while trimming barcode and restriction enzyme sequences from retained reads. Afterwards, the genetic variants were discovered using Genome Analysis Toolkit (GATK) v4.2.6.1 (McKenna et al. 2010), according to the GATK Best Practices workflows (Depristo et al. 2011; Van der Auwera and O'Connor 2020). The reads were aligned to the reference genome Williams 82.a5 (Garg et al. 2023), using the Burrows-Wheeler Aligner (BWA) software (Li 2013). The HaplotypeCaller was used to call individual genetic variants, and joint genotyping of all samples was subsequently performed with GenotypeGVCFs to generate a unified SNP dataset.

Data analysis

To retain only high-confidence variants, SNPs were filtered using GATK VariantFiltration with the criteria QUAL > 30 and DP > 10. The quality threshold (QUAL > 30) ensured a variant-calling accuracy of more than 99.9%, reducing false-positive calls arising from sequencing errors. The depth of coverage threshold (DP > 10) guarantees that each retained variant was supported by at least 10 sequencing reads, providing reliable genotyping confidence. The high-quality SNP dataset was imputed using the LD KNNi method and subsequently filtered in TASSEL v.5.2.96 (Bradbury et al. 2007; Glaubitz et al. 2014). Imputation was performed to infer all the missing genotypes while filtering retained only loci with a genotype call rate (GCR) ≥ 90%, and minor allele frequency (MAF) ≥ 5%. The marker density and the genome-wide SNP distribution were visualized using the R 4.4.2 rMVP package (Yin et al. 2021).

Population structure

The genetic structure of the soybean population was assessed using multiple approaches. Genetic relationships among genotypes were first explored using principal component analysis (PCA) implemented in the R 4.4.2 adegenet package (Jombart 2008). Population structure was then inferred

using the LEA package in R version 4.4.2 by testing K values from 1 to 10, with the optimal number of clusters determined based on the cross-entropy criterion (Frichot and François 2015). While examining the cross-entropy, three levels of genetic structure were examined: K = 2 to assess major biological differentiation, K = 3 to identify the most likely structure, and K = 5 to explore sub-clustering. Individuals were assigned to genetic clusters based on an ancestry coefficient (Q) ≥ 0.7, whereas admixed individuals (Q < 0.7) were excluded from subsequent genetic diversity analyses to minimize the confounding effects of mixed ancestry and ensure that diversity estimates reflected genetically coherent subpopulations (Frichot et al. 2014). To further validate population structure, discriminant analysis of principal components (DAPC), was performed in adegenet (Jombart and Ahmed 2011), to maximize between-cluster variation and minimize within-cluster variation (Jombart et al. 2010). The reconstruction of phylogenetic relationships in the population was also carried out using the neighbour-joining method with the phangorn package (Schliep et al. 2017) to generate a Newick (nwk) tree, which was further refined in MEGA v12 (Kumar et al. 2024).

Population differentiation

The analysis of molecular variance (AMOVA) was performed using the R poppr package to partition total genetic variation among and within genetic groups (Kamvar et al. 2014). To assess the degree of divergence among subpopulations, pairwise fixation index (F_{ST}) was computed with the R adegenet and poppr packages (Jombart and Ahmed 2011; Kamvar et al. 2014). Moreover, pairwise Nei's genetic distances and genetic identities among subpopulations were computed using the same packages to quantify genetic differentiation between clusters.

Genetic diversity analyses

Genetic diversity within the inferred subpopulations was estimated using the SnpReady package in R version 4.4.2 (Granato et al. 2018). The parameters calculated included polymorphic information content (PIC), minor allele frequency (MAF), expected heterozygosity (He), observed heterozygosity (Ho), fixation index (F_{IS}), effective number of alleles (Ne), additive genetic variance (Va), and dominance variance (Vd).

Results

Genome-wide SNP distribution across the soybean genome

RAD sequencing of 191 soybean genotypes from diverse origins yielded 86,592 raw SNPs, of which 21,970 high-quality markers (25.37%) were retained after imputation and filtering (GCR ≥ 90%, MAF ≥ 5%). Genome-wide SNPs were distributed across all 20 soybean chromosomes, although their distribution was highly heterogeneous (Fig. 2). SNP density varied markedly among genomic regions, with most 1-Mb windows exhibiting low to moderate marker densities and a limited number of genomic intervals

displaying pronounced SNP enrichment. The most prominent hotspot was located on chromosome 18, where a maximum of 9,597 SNPs were detected within a single 1-Mb window. Additional regions of elevated SNP density were observed on chromosomes 4, 8, 10, and 17, whereas several chromosomal regions contained comparatively few polymorphic sites. The broad genomic coverage and heterogeneous SNP distribution demonstrate that the marker dataset captures genome-wide sequence variation and provides sufficient resolution for subsequent analyses of population structure, genetic diversity, and phylogenetic relationships.

Population structure reveals three major genetic groups

Population structure analysis revealed a sharp decline in cross-entropy values from K = 1 to K = 3, followed by a gradual decrease with increasing values of K, indicating that the major genetic structure of the soybean panel was captured by three ancestral populations (Fig. 3a). At K = 2, the genotypes were separated into two major genetic clusters that largely corresponded to vegetable soybeans and grain soybeans, although substantial admixture was observed between the two germplasm types (Fig. 3b). Increasing the number of clusters to K = 3 further resolved the panel into three distinct genetic clusters, while maintaining clear assignment of most accessions (Fig. 3c). Further subdivision at K = 5 partitioned the three major groups into additional smaller clusters, three of which contained fewer than 10 genotypes, providing limited additional biological resolution (Fig. 3d). While using an ancestry coefficient threshold of Q ≥ 0.7, 116 (60.73%), 121 (63.35%), and 95 (49.74%) accessions were assigned to a single ancestral population at K = 2, K = 3, and K = 5, respectively. The highest proportion of well-assigned accessions at K = 3, together with its clear representation of the major genetic groups, supported its selection as the most biologically informative level of population subdivision for subsequent analyses (Table 1).

Table 1
Statistics of pure genotypes among this soybean collection at K = 2, 3, and 5 with Q ≥ 0.7

Number of ancestral populations (K)	Total number of pure genotypes per ancestral population	Pure genotypes per cluster		Percentage (%) of vegetable soybean within cluster
		Cluster	**size**	
2	116	1	21	100
		2	95	49.5
3	121	1	34	100
		2	5	100
		3	82	46.3
5	95	1	7	100
		2	12	91.7
		3	5	100
		4	3	0
		5	68	55.9

Genetic differentiation between grain and vegetable soybean accessions

Principal component analysis (PCA) of this soybean collection revealed partial genetic differentiation between grain and vegetable soybean accessions (Fig. 4a). The first two components explained 5.64% (PC1) and 2.52% (PC2) of the total variation, respectively. Although considerable overlap was observed between the two soybean types, grain soybean accessions formed a relatively compact cluster predominantly distributed along the negative side of PC1. In contrast, vegetable soybean accessions exhibited a much broader distribution across both principal components, particularly extending toward the positive end of PC1, indicating greater genetic heterogeneity. The extensive overlap between the two groups suggests a shared genetic background, whereas the narrower dispersion of grain soybean accessions indicates reduced genetic diversity, likely reflecting the use of a relatively narrow genetic base during soybean improvement.

Concordant multivariate and phylogenetic analyses validate the inferred population structure

Principal component analysis (PCA) based on the three inferred genetic clusters (K = 3) revealed clear genetic differentiation among the soybean accessions (Fig. 4b). Cluster 1 (green) and Cluster 2 (orange), consisting predominantly of vegetable soybean accessions, were separated primarily along PC1, whereas Cluster 3 (purple), which contained a higher proportion of grain soybean accessions, was

clearly differentiated by both PC1 and PC2. A small number of admixed accessions occupied intermediate positions between the major clusters, reflecting shared ancestry among populations. Discriminant analysis of principal components (DAPC) further resolved three inferred populations into well-defined genetic groups with minimal overlap, indicating strong genetic differentiation among clusters (Fig. 4c). The first two discriminant functions captured the majority of the between-group variation, producing compact clusters while maintaining the relationships observed in the PCA. Consistent with these results, the neighbor-joining phylogenetic tree grouped the accessions into three major clades that closely corresponded to the clusters identified by both PCA and DAPC (Fig. 4d). Clusters 1 and 2 were composed primarily of vegetable soybean accessions, whereas Cluster 3 consisted predominantly of grain soybean accessions. The high concordance among PCA, DAPC, and phylogenetic analyses confirms the presence of three major genetic groups within the soybean core collection and supports the population structure.

Genetic differentiation among inferred populations

Analysis of molecular variance (AMOVA) of the pure soybean population indicated that 83.78% of the total genetic variation was distributed within clusters, whereas 16.22% was attributed to differences among clusters (Table 2). The among-cluster component was highly significant ($\Phi = 0.162$, $p = 0.001$), indicating statistically significant genetic differentiation among the three inferred populations. The predominance of within-cluster variation suggests that substantial genetic diversity is maintained within each population despite their overall differentiation.

Table 2

Analysis of molecular variance (AMOVA)[a] among the genetic clusters from population structure analysis

Source of Variation	df	Sum of Squares	Mean Square	Variance Component	% Total Variation	[b]Phi-statistic	[c]p-value
Among subpopulations	2	55680.54	27840.27	843.61	16.22	0.162	0.001***
Within subpopulations	118	514333.61	4358.76	4358.76	83.78		
Total	120	570014.15	4750.12	5202.3648	100		

[a]AMOVA carried out using 999 permutations

[b]Phi-statistic corresponds to Φ_{ST} (Phi-samples-total)

[c]p-value obtained from the Monte-Carlo permutation test

Pairwise fixation indices (F_{ST}) further quantified the degree of genetic differentiation among clusters (Table S2). The greatest differentiation was observed between clusters 2 and 3 ($F_{ST} = 0.335$), indicating strong genetic divergence. Clusters 1 and 2 also showed moderate to high differentiation ($F_{ST} = 0.176$).

The lowest differentiation was detected between clusters 1 and 3 (F_{ST} = 0.128), suggesting a comparatively closer genetic relationship between those groups.

Nei's genetic identity and genetic distance analyses were consistent with the F_{ST} estimates (Table S3). The greatest genetic distance was observed between clusters 2 and 3 (D = 0.145; I = 0.865), followed by clusters 1 and 2 (D = 0.115; I = 0.891). In contrast, clusters 1 and 3 exhibited the smallest genetic distance (D = 0.043) and the highest genetic identity (I = 0.958), suggesting a closer evolutionary relationship and a greater shared ancestry. Collectively, the AMOVA, F_{ST}, and Nei's genetic distance analyses consistently demonstrated moderate but significant genetic differentiation among the three populations, corroborating the clustering patterns identified by the population structure, PCA, DAPC, and phylogenetic analyses.

Genetic diversity of the inferred populations

Genetic diversity parameters varied across the three inferred clusters (Table 3). Cluster 1 exhibited the highest levels of genetic diversity, with the greatest polymorphic information content (PIC = 0.253), minor allele frequency (MAF = 0.172), expected heterozygosity (He = 0.262), and effective number of alleles (Ne = 1.393), indicating greater allelic richness within this group. In contrast, Cluster 3 showed the lowest values for these diversity indices (PIC = 0.188, MAF = 0.119, He = 0.178, Ne = 1.268), suggesting a comparatively narrower genetic base despite its larger sample size. Cluster 2, although comprising only five accessions, displayed intermediate levels of genetic diversity (He = 0.246) together with the highest additive genetic variance (Va = 0.086) and dominance variance (Vd = 21.924).

Table 3
Genetic diversity within the genetic clusters

		Diversity parameters							
Cluster	**Individuals**	**PIC**	**MAF**	**He**	**Ho**	**F_{IS}**	**Ne**	**Va**	**Vd**
Cluster 1	34	0.253	0.172	0.262	0.013	0.949	1.393	0.027	18.016
Cluster 2	5	0.193	0.143	0.246	0.019	0.925	1.333	0.086	21.924
Cluster 3	82	0.188	0.119	0.178	0.009	0.952	1.268	0.021	12.262
Total	121	0.246	0.163	0.228	0.014	0.941	1.370	0.018	14.444

Observed heterozygosity (Ho) was low across all three clusters (0.009–0.019), where fixation indices (F_{IS}) were uniformly high (0.925–0.952), reflecting the highly homozygous nature of cultivated soybean. Across the entire population, intermediate diversity values were observed (PIC = 0.246, He = 0.228), indicating moderate overall genetic diversity, despite the clear genetic differentiation among clusters. The contrasting diversity patterns among clusters indicate that the inferred populations differ not only in genetic composition but also in the extent of allelic variation retained within each group. In particular, the relatively high diversity observed in Cluster 1 and the comparatively narrow genetic base of Cluster 3 are

consistent with the population structure and phylogenetic analyses, highlighting unequal distribution of genetic variation within the soybean core collection.

Discussion

Population structure reflects domestication, breeding, and geographic adaptation

Genome-wide SNP analysis consistently resolved this soybean panel into three major genetic groups, as supported by the concordance among population structure, PCA, DAPC, and phylogenetic analyses. Although the first two principal components explained only a modest proportion of the total molecular variation, the consistent clustering obtained across multiple analytical methods indicates that the inferred population structure is robust. Similar studies using genome-wide SNP markers have identified two to six major genetic groups in cultivated soybean, reflecting differences in domestication history, breeding objectives, and geographic origin (Liu et al. 2017; Shaibu et al. 2022; Silue et al. 2025).

The observed population structure reflects the combined effects of domestication, crop improvement, and germplasm exchange (Nguyen and Bhattacharyya 2017; Jeong et al. 2019). Following domestication from *Glycine soja*, cultivated soybean underwent strong artificial selection for traits including reduced pod shattering, increased seed size, flowering time, and adaptation to diverse agroecological environments (Hyten et al. 2006; Li et al. 2013; Sedivy et al. 2017). Subsequent dissemination across different geographic regions and recurrent use of elite breeding lines further reshaped allele frequencies while maintaining substantial shared ancestry among cultivated soybean populations (Valliyodan et al. 2016, 2019). Accordingly, the admixed accessions identified in this study likely represent historical hybridization and germplasm exchange among breeding programs and geographic regions (Li et al. 2022; Valliyodan et al. 2016). The three genetic groups also showed a close association with soybean market class. Clusters 1 and 2 consisted predominantly of vegetable soybean accessions, whereas Cluster 3 contained mainly grain soybean accessions. Similar differentiation between grain and vegetable soybean has been reported previously and is largely attributed to contrasting breeding objectives (Liu et al. 2022, 2025a). Vegetable soybean breeding has focused primarily on fresh pod quality, large seed size, sweetness, texture, and consumer preference (Nair et al. 2022; Agyenim-Boateng et al. 2023; Meghana et al. 2025), whereas grain soybean improvement has emphasized grain yield, lodging resistance, disease resistance, abiotic stress tolerance, and seed composition (Guo et al. 2022; Nawaz et al. 2026). These contrasting selection pressures have gradually altered allele frequencies, resulting in the distinct population structure observed in cultivated soybean

(Dong et al. 2014; Liu et al. 2025b). Geographic origin further contributed to the observed clustering pattern. Most African grain soybean cultivars were grouped within Cluster 3, whereas vegetable soybean accessions originating mainly from East Asia and North America were distributed across Clusters 1 and 2. Soybean cultivation in Sub-Saharan Africa is relatively recent (Carter et al. 2004; Sedivy et al. 2017). It has relied heavily on introduced germplasm from Asia, North America, and international breeding

programs coordinated by IITA and collaborating institutions (Chigeza et al. 2019; Khojely et al. 2018; Tukamuhabwa and Oloka 2016). The clustering of African cultivars therefore suggests that current breeding populations originated from a relatively narrow pool of founder lines that have been repeatedly used during cultivar development (Hyten et al. 2006; Valliyodan et al. 2016).

Genetic diversity provides evidence for broadening the African soybean gene pool

Although three distinct populations were identified, AMOVA demonstrated that nearly 84% of the total genetic variation resided within populations, while only 16% occurred among populations. The predominance of diversity within clusters revealed the availability of significant allelic variation for selection in each subpopulation (Shaibu et al. 2022; Silue et al. 2025; Yao et al. 2023). This pattern agrees with previous studies in cultivated soybean and other self-pollinated crops, where most genetic diversity is retained within breeding populations despite significant population differentiation (Assfaw et al. 2025; Lukanda et al. 2023; Shaibu et al. 2022; Silue et al. 2025). The moderate F_{ST} values and Nei's genetic distances further indicate that the three populations have diverged genetically while maintaining substantial shared ancestry. This differentiation across subpopulations shows the breeding value of genetically diverse clusters, as divergent germplasm boosts the potential of recombination and enables the introduction of novel genes associated with agronomic and nutritional traits (Viana et al. 2022; Niu et al. 2024)

The distribution of genetic diversity differed considerably among the three populations. Cluster 1 exhibited the highest polymorphic information content, expected heterozygosity, and effective number of alleles, indicating that it retains the broadest allelic diversity within the collection. Conversely, Cluster 3 displayed consistently lower diversity despite containing most African grain soybean cultivars. This reduced diversity is likely attributable to repeated selection from a limited number of elite founder lines during soybean improvement in Sub-Saharan Africa. Likewise, reductions in diversity have been widely reported following domestication and modern breeding, where intensive selection narrows the genetic base of elite cultivars (Jeong et al. 2019; Chigeza et al. 2019; Lukanda et al. 2023; Wang et al. 2019). The contrasting diversity observed among the three populations has important implications for germplasm conservation. The relatively high diversity maintained within Clusters 1 and 2 indicates that these vegetable soybean accessions constitute an important reservoir of allelic variation that remains underutilized in African soybean breeding. In contrast, the comparatively narrow diversity of Cluster 3 highlights the need to broaden the genetic base of African breeding populations through strategic utilization of genetically diverse germplasm (Hyten et al. 2006; Zhuang et al. 2022).

Implications for vegetable soybean breeding in Sub-Saharan Africa

Vegetable soybean remains an underutilized crop in Sub-Saharan Africa (Williams and Haq 2002; Geetha et al. 2023; Ndlovu et al. 2024), despite its recognized nutritional value and potential contribution to diversified food systems (Agyenim-Boateng et al. 2023; Soujanya et al. 2025; Meghana et al. 2025).

The present study demonstrates that the introduction of the World Vegetable Center large-seeded soybean core collection represents a valuable genetic resource for addressing this challenge. The collection contains substantial genome-wide diversity distributed two major genetic groups while retaining close genetic relationships with African grain soybean cultivars. This relationship suggests that favorable alleles controlling large seed size, pod quality, and consumer-preferred traits can be incorporated into locally adapted African breeding materials with relatively limited genetic incompatibility.

From a breeding perspective, the identification of genetically distinct populations provides a practical framework for parental selection. Crosses involving genetically divergent clusters are expected to increase allelic recombination and facilitate the incorporation of novel alleles for both agronomic and quality traits. Meanwhile, accessions within the same cluster may serve as valuable materials for improving specific target traits while maintaining adaptation to regional production environments. The genome-wide SNP dataset generated in this study also establishes a genomic foundation for future marker-assisted selection, genomic selection, and the development of core breeding populations for vegetable soybean improvement.

Overall, this study represents the first comprehensive genome-wide characterization of the vegetable soybean genetic resources. The identified population structure and genetic diversity provide valuable information for germplasm conservation, efficient parental selection, and the strategic utilization of vegetable soybean genetic resources. These findings will facilitate the development of nutritionally superior, large-seeded vegetable soybean cultivars adapted to Sub-Saharan Africa and contribute to the long-term diversification and resilience of soybean breeding programs in the region.

Conclusion

This study provides a comprehensive genomic background for optimizing the soybean germplasm utilization in breeding and improvement programs. Our multi-analysis approach successfully resolved the genetic structure of the diverse soybean germplasm collection into three main genetic groups, driven largely by breeding history, market preference, and environmental adaptation over time. The clear separation between the vegetable and grain soybean, coupled with the identification of pure lines among the germplasm, offers an opportunity for targeted parental selection in future breeding programs. Furthermore, while the low observed heterozygosity confirms the expected self-pollinating nature of this collection, the moderate genetic diversity and evidence of historical gene flow highlight a valuable reservoir of alleles. Overall, our findings map out the genetic diversity of a global collection of soybean genotypes, offering a clear roadmap for future vegetable soybean breeding.

Declarations

Consent for publication

All authors have read and approved the final manuscript and give their explicit consent for its publication.

Competing interest

The authors have no competing interests.

Funding

This research was mainly supported by the Partnership for Applied Skills in Sciences, Engineering, and Technology-Regional Scholarship and Innovation Fund [(PASET-RSIF), grant number: B8501G30190] and the Carnegie Corporation of New York.

Author Contribution

M.K.A.D. and Y.-P.L. conceived and designed the study. M.K.A.D. and Y.-P.L. developed the methodology. M.K.A.D. and E.O.O. performed the formal analyses and curated the data. M.K.A.D. and Y.-P.L. wrote the original draft of the manuscript. E.E.A., E.O.O., P.T., T.O., M.O.-S., E.N., I.O.D., P.G., T.L.O., F.J.-B.Q., R.M.N., R.E., A.A., and Y.-P.L. reviewed and edited the manuscript. Y.-P.L., E.E.A., and P.T. supervised the research. All authors read and approved the final manuscript.

Acknowledgement

The authors acknowledge the financial support of the Partnership for Applied Skills in Sciences, Engineering, and Technology-Regional Scholarship and Innovation Fund (PASET-RSIF) and the Carnegie Corporation of New York. We sincerely thank the Seed Savers Exchange (USA) and World Vegetable Center (Taiwan) for sharing genetic materials. We further express our gratitude to the World Vegetable Center's long-term strategic donors, including Taiwan, the United States, Australia, the United Kingdom, Germany, Thailand, South Korea, the Philippines, and Japan.

Data Availability

The genetic sequencing datasets generated and analyzed during the current study are available in the NCBI Sequence Read Archive repository under the BioProject number PRJNA1454238.

References

1. Agoyi EE, Ahomondji SE, Yemadje PL, et al (2024) Combining AMMI and BLUP analysis to select high-yielding soybean genotypes in Benin. Agron J 116:2109–2128. https://doi.org/10.1002/agj2.21615

2. Agyenim-Boateng KG, Zhang S, Zhang S, et al (2023) The nutritional composition of the vegetable soybean (maodou) and its potential in combatting malnutrition. Front Nutr 9:1–21. https://doi.org/10.3389/fnut.2022.1034115
3. Ahomondji ES, Agoyi EE, Lokonon BE, et al (2022) Multi-Environmental Evaluation of Vegetable Soybean for Adaptation and Stability in Benin. Afr Crop Sci J 30:313–330. https://doi.org/10.4314/acsj.v30i3.4
4. Araújo MS, Chaves S, Ferreira GNC, et al (2025) High-resolution soybean trial data supporting the expansion of agriculture in Africa. Sci Data 12:1908. https://doi.org/10.1038/s41597-025-06190-3
5. Assfaw AT, Bunmi O, Paterne A, et al (2025) Genetic diversity and population structure analysis of soybean [Glycine max (L.) Merrill] genotypes based on agro-morphological traits and SNP markers. PLOS One 20:e0332895. https://doi.org/10.1371/journal.pone.0332895
6. Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308
7. Carter TE, Nelson RL, Sneller CH, Cui Z (2004) Genetic Diversity in Soybean. In: Shibles RM, Harper JE, Wilson RF, Shoemaker RC (eds) Agronomy Monographs, 1st edn. Wiley, pp 303–416
8. Catchen J, Hohenlohe PA, Bassham S, et al (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140. https://doi.org/10.1111/mec.12354
9. Chigeza G, Boahen S, Gedil M, et al (2019) Public sector soybean (Glycine max) breeding: Advances in cultivar development in the African tropics. Plant Breed 138:455–464. https://doi.org/10.1111/pbr.12682
10. Depristo MA, Banks E, Poplin R, et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data
11. Dong L, Fang C, Cheng Q, et al (2021) Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. Nat Commun 12:5445. https://doi.org/10.1038/s41467-021-25800-3
12. Dong Y, Yang X, Liu J, et al (2014) Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nat Commun 5:3352. https://doi.org/10.1038/ncomms4352
13. Elshire RJ, Glaubitz JC, Sun Q, et al (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6:1–10. https://doi.org/10.1371/journal.pone.0019379
14. Fang C, Du H, Wang L, et al (2024) Mechanisms underlying key agronomic traits and implications for molecular breeding in soybean. J Genet Genomics 51:379–393. https://doi.org/10.1016/j.jgg.2023.09.004
15. Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. Methods Ecol Evol 6:925–929. https://doi.org/10.1111/2041-210X.12382
16. Frichot E, Mathieu F, Trouillon T, et al (2014) Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics 196:973–983. https://doi.org/10.1534/genetics.113.160572

17. Garg V, Khan AW, Fengler K, et al (2023) Near-gapless genome assemblies of Williams 82 and Lee cultivars for accelerating global soybean research. Plant Genome 16:e20382. https://doi.org/10.1002/tpg2.20382
18. Geetha A, Pidigam A, Suresh JD, Raviteja R (2023) Underutilized vegetables crops for improved nutrition and health: A review. J Phytonanotechnology Pharm Sci 3:. https://doi.org/10.54085/jpps.2023.3.4.2
19. Glaubitz JC, Casstevens TM, Lu F, et al (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE 9:e90346. https://doi.org/10.1371/journal.pone.0090346
20. Granato ISC, Galli G, De Oliveira Couto EG, et al (2018) snpReady: a tool to assist breeders in genomic analysis. Mol Breed 38:102. https://doi.org/10.1007/s11032-018-0844-8
21. Gueco LS, Tejano MS, Yanos LA, et al (2022) Conservation of Musa Diversity and Germplasm Management at the National Plant Genetic Resources Laboratory in the Philippines. In: Ramamoorthy S, Buot IJr, Chandrasekaran R (eds) Plant Genetic Resources, Inventory, Collection and Conservation. Springer, Singapore, pp 242–258
22. Guo B, Sun L, Jiang S, et al (2022) Soybean genetic resources contributing to sustainable protein production. Theor Appl Genet 135:4095–4121. https://doi.org/10.1007/s00122-022-04222-9
23. Hartman GL, West ED, Herman TK (2011) Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. Food Secur 3:5–17. https://doi.org/10.1007/s12571-010-0108-x
24. He S-L, Wang Y-S, Li D-Z, Yi T-S (2016) Environmental and Historical Determinants of Patterns of Genetic Differentiation in Wild Soybean (Glycine soja Sieb. et Zucc). Sci Rep 6:22795. https://doi.org/10.1038/srep22795
25. Huang Y, Ku H, Wang C, et al (2022) A multiple phenotype imputation method for genetic diversity and core collection in Taiwanese vegetable soybean. Front Plant Sci 13:1–20. https://doi.org/10.3389/fpls.2022.948349
26. Hyten DL, Song Q, Zhu Y, et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci 103:16666–16671. https://doi.org/10.1073/pnas.0604379103
27. Jeong S-C, Moon J-K, Park S-K, et al (2019) Genetic diversity patterns and domestication origin of soybean. Theor Appl Genet 132:1179–1193. https://doi.org/10.1007/s00122-018-3271-7
28. Jiang B, Nan H, Gao Y, et al (2014) Allelic Combinations of Soybean Maturity Loci E1, E2, E3 and E4 Result in Diversity of Maturity and Adaptation to Different Latitudes. PLoS ONE 9:e106042. https://doi.org/10.1371/journal.pone.0106042
29. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129
30. Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071. https://doi.org/10.1093/bioinformatics/btr521
31. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94.

https://doi.org/10.1186/1471-2156-11-94

32. Kamvar ZN, Tabima JF, Grünwald NJ (2014) *Poppr*: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281. https://doi.org/10.7717/peerj.281
33. Khojely DM, Ibrahim SE, Sapey E, Han T (2018) History, current status, and prospects of soybean production and research in sub-Saharan Africa. Crop J 6:226–235. https://doi.org/10.1016/j.cj.2018.03.006
34. Kumar S, Stecher G, Suleski M, et al (2024) MEGA12: Molecular Evolutionary Genetic Analysis Version 12 for Adaptive and Green Computing. 41:1–9. https://doi.org/https://doi.org/10.1093/molbev/msae263
35. Li C, Wang W, Pan Y, et al (2022) Germplasm Sources, Genetic Richness, and Population Differentiation of Modern Chinese Soybean Cultivars Based on Pedigree Integrated With Genomic-Marker Analysis. Front Plant Sci 13:945839. https://doi.org/10.3389/fpls.2022.945839
36. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997
37. Li Y, Li W, Zhang C, et al (2010) Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytol 188:242–253. https://doi.org/10.1111/j.1469-8137.2010.03344.x
38. Li Y, Zhao S, Ma J, et al (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14:579. https://doi.org/10.1186/1471-2164-14-579
39. Liu N, Feng W, Zhang G, et al (2025a) Genomic and population evidence uncovers divergent improvement of vegetable soybean from grain soybean. Mol Plant 18:1094–1097. https://doi.org/10.1016/j.molp.2025.06.009
40. Liu N, Feng W, Zhang G, et al (2025b) Genomic and population evidence uncovers divergent improvement of vegetable soybean from grain soybean. Mol Plant 18:1094–1097. https://doi.org/10.1016/j.molp.2025.06.009
41. Liu N, Li M, Hu X, et al (2017) Construction of high-density genetic map and QTL mapping of yield-related and two quality traits in soybean RILs population by RAD-sequencing. BMC Genomics 18:1–13. https://doi.org/DOI%2010.1186/s12864-017-3854-8
42. Liu N, Niu Y, Zhang G, et al (2022) Genome sequencing and population resequencing provide insights into the genetic basis of domestication. Hortic Res 9:1–12. https://doi.org/10.1093/hr/uhab052
43. Lu S, Zhao X, Hu Y, et al (2017) Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. Nat Genet 49:773–779. https://doi.org/10.1038/ng.3819
44. Lukanda MM, Dramadri IO, Adjei EA, et al (2023) Genetic Diversity and Population Structure of Ugandan Soybean (Glycine max L.) Germplasm Based on DArTseq. Plant Mol Biol Report 1–20. https://doi.org/10.1007/s11105-023-01375-9

45. Maldonado Dos Santos JV, Sant'Ana GC, Wysmierski PT, et al (2022) Genetic relationships and genome selection signatures between soybean cultivars from Brazil and United States after decades of breeding. Sci Rep 12:10663. https://doi.org/10.1038/s41598-022-15022-y
46. McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110
47. Meghana D, Rachana KS, Pan RS, et al (2025) Nutritional analysis of vegetable soybean [Glycine max (L.) Merrill] accessions in Eastern India. Front Nutr 12:1643470. https://doi.org/10.3389/fnut.2025.1643470
48. Moinuddin, Khan FA, Amir M, et al (2023) Vegetable soybean (Edamame): A potential area of research - A review. SKUAST J Res 25:376–386. https://doi.org/10.5958/2349-297x.2023.00051.x
49. Moseley D, Mozzoni L, Kaler A, et al (2021) Evaluation of genetic diversity and association mapping for seed weight and size in vegetable soybean germplasm. Crop Sci 61:3516–3528. https://doi.org/10.1002/csc2.20588
50. Myles S, Peiffer J, Brown PJ, et al (2009) Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. Plant Cell 21:2194–2202. https://doi.org/10.1105/tpc.109.068437
51. Nair RM, Boddepalli VN, Yan MR, et al (2023) Global Status of Vegetable Soybean. Plants 12:. https://doi.org/10.3390/plants12030609
52. Nair RM, Yan M rong, Vemula AK, et al (2022) Development of core collections in soybean on the basis of seed size. Legume Sci 5:1–6. https://doi.org/10.1002/leg3.158
53. Nawaz MA, Chung G, Pamirsky IE, Golokhvast KS (2026) Breeding Climate-Resilient Soybeans for 2050 and Beyond: Leveraging Novel Technologies to Mitigate Yield Stagnation and Climate Change Impacts. Plants 15:1201. https://doi.org/10.3390/plants15081201
54. Ndlovu M, Scheelbeek P, Ngidi M, Mabhaudhi T (2024) Underutilized crops for diverse, resilient and healthy agri-food systems: a systematic review of sub-Saharan Africa. Front Sustain Food Syst 8:1498402. https://doi.org/10.3389/fsufs.2024.1498402
55. Nguyen HT, Bhattacharyya MK (2017) The Soybean Genome. Springer International Publishing AG 2017
56. Niu Y, Yung W-S, Sze C-C, et al (2024) Developing an SNP dataset for efficiently evaluating soybean germplasm resources using the genome sequencing data of 3,661 soybean accessions. BMC Genomics 25:475. https://doi.org/10.1186/s12864-024-10382-3
57. Powell W, Morgante M, Andre C, et al (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238. https://doi.org/10.1007/BF00564200
58. Rochette NC, Rivera-Colón AG, Catchen JM (2019) Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol 28:4737–4754. https://doi.org/10.1111/mec.15253

59. Schliep K, Potts AJ, Morrison DA, Grimm GW (2017) Intertwining phylogenetic trees and networks. Methods Ecol Evol 8:1212–1220. https://doi.org/10.1111/2041-210X.12760
60. Sedivy EJ, Wu F, Hanzawa Y (2017) Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol 214:539–553. https://doi.org/10.1111/nph.14418
61. Shaibu AS, Ibrahim H, Miko ZL, et al (2022) Assessment of the Genetic Structure and Diversity of Soybean Using Diversity Array Technology and Single Nucleotide Polymorphism Markers. Plants 11:1–13. https://doi.org/10.3390/plants11010068
62. Silue T, Agre PA, Olasanmi B, et al (2025) Genetic diversity and population structure of soybean (Glycine max (L.) Merril) germplasm. https://doi.org/10.1101/2024.10.02.616345
63. Song Q, Hyten DL, Jia G, et al (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE 8:e54985. https://doi.org/10.1371/journal.pone.0054985
64. Soujanya K, Supraja T, Kuna A, et al (2025) Comparative Evaluation of Nutritional Quality and In Vitro Protein Digestibility in Selected Vegetable Soybean Genotypes at R6 and R8 Maturity. Foods 14:1–29. https://doi.org/10.3390/foods14142549
65. Tsindi A, Kawuki R, Tukamuhabwa P (2019) Adaptation and Stability Of Vegetable Soybean Genotypes. Afr Crop Sci J 27:267–280
66. Tukamuhabwa P, Oloka HK (2016) Soybean Research & Development in Uganda: A case of paradigm shift in an African University. Makerere University Agricultural Research Institute, Kabanyolo (MUARIK), Center for Soybean Improvement and Development, Makerere University, Kampala
67. Valliyodan B, Brown AV, Wang J, et al (2021) Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. Sci Data 8:50. https://doi.org/10.1038/s41597-021-00834-w
68. Valliyodan B, Cannon SB, Bayer PE, et al (2019) Construction and comparison of three reference-quality genome assemblies for soybean. Plant J 100:1066–1082. https://doi.org/10.1111/tpj.14500
69. Valliyodan B, Dan Qiu, Patil G, et al (2016) Landscape of genomic diversity and trait discovery in soybean. Sci Rep 6:23598. https://doi.org/10.1038/srep23598
70. Van der Auwera GA, O'Connor BD (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, 1st edn. O'Reilly Media
71. Van Der Merwe R, Labuschagne MT, Smit A (2024) Cultivar variability and stability of vegetable-type soybean for seed yield and pod shattering. South Afr J Bot 166:106–115. https://doi.org/10.1016/j.sajb.2024.01.034
72. Viana JPG, Avalos A, Zhang Z, et al (2024) Common signatures of selection reveal target loci for breeding across soybean populations. Plant Genome 17:e20426. https://doi.org/10.1002/tpg2.20426
73. Viana JPG, Fang Y, Avalos A, et al (2022) Impact of multiple selective breeding programs on genetic diversity in soybean germplasm. Theor Appl Genet 135:1591–1602. https://doi.org/10.1007/s00122-022-04056-5

74. Wang J, Chu S, Zhang H, et al (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6:20728. https://doi.org/10.1038/srep20728
75. Wang X, Chen L, Ma J (2019) Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. Genome Biol 20:22. https://doi.org/10.1186/s13059-019-1631-5
76. Williams JT, Haq N (2002) Global research on underutilized crops. An assessment of current activities and proposals for enhanced cooperation. International Centre for Underutilised Crops, Southampton, UK
77. Yao X, Xu J yuan, Liu Z xiong, et al (2023) Genetic diversity in early maturity Chinese and European elite soybeans: A comparative analysis. Euphytica 219:1–14. https://doi.org/10.1007/s10681-022-03147-0
78. Yin L, Zhang H, Tang Z, et al (2021) rMVP: A Memory-Efficient, Visualization-Enhanced, and Parallel-Accelerated Tool for Genome-Wide Association Study. Genomics Proteomics Bioinformatics 19:619–628. https://doi.org/10.1016/j.gpb.2020.10.007
79. Yu X, Fu X, Yang Q, et al (2022) Genome-Wide Variation Analysis of Four Vegetable Soybean Cultivars Based on Re-Sequencing. Plants 11:28. https://doi.org/10.3390/plants11010028
80. Zhou Z, Jiang Y, Wang Z, et al (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:408–414. https://doi.org/10.1038/nbt.3096
81. Zhuang Y, Li X, Hu J, et al (2022) Expanding the gene pool for soybean improvement with its wild relatives. aBIOTECH 3:115–125. https://doi.org/10.1007/s42994-022-00072-7

Figures

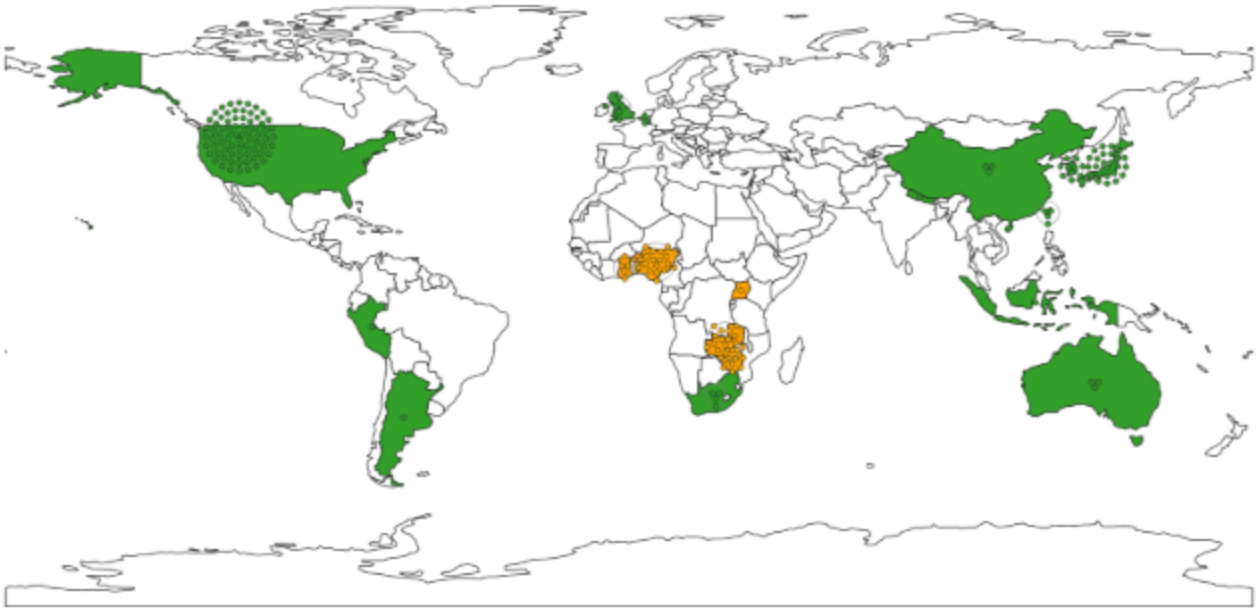

Figure 1

Distribution of the 191 soybean accessions used in this study based on passport data. Green locations depict sources of vegetable soybeans, and orange locations show sources of grain soybeans. Likewise, green dots represent the number of vegetable soybeans, while orange dots represent the number of grain soybeans obtained from each location.

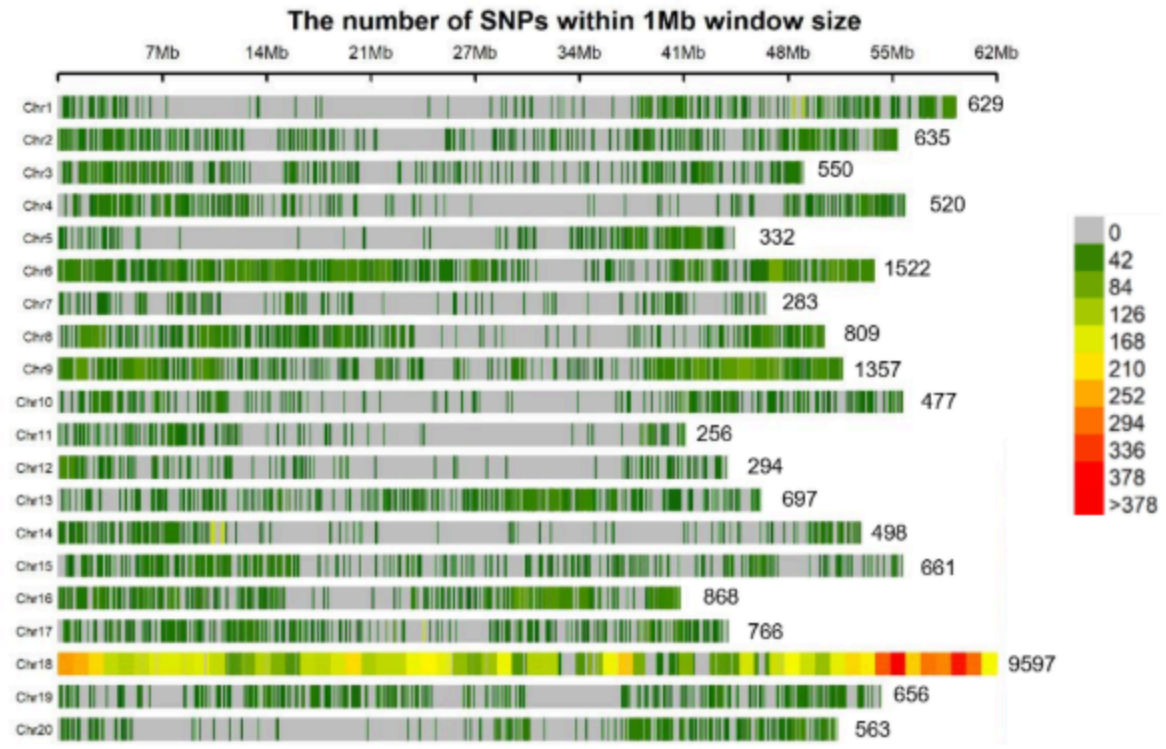

Figure 2

Genome-wide distribution of SNP density across the 20 soybean chromosomes. SNP density was calculated using non-overlapping 1-Mb genomic windows. Colors represent the number of SNPs within each window according to the scale shown on the right, with redder colors indicating higher SNP densities. Numbers at the end of each chromosome indicate the number of SNPs.

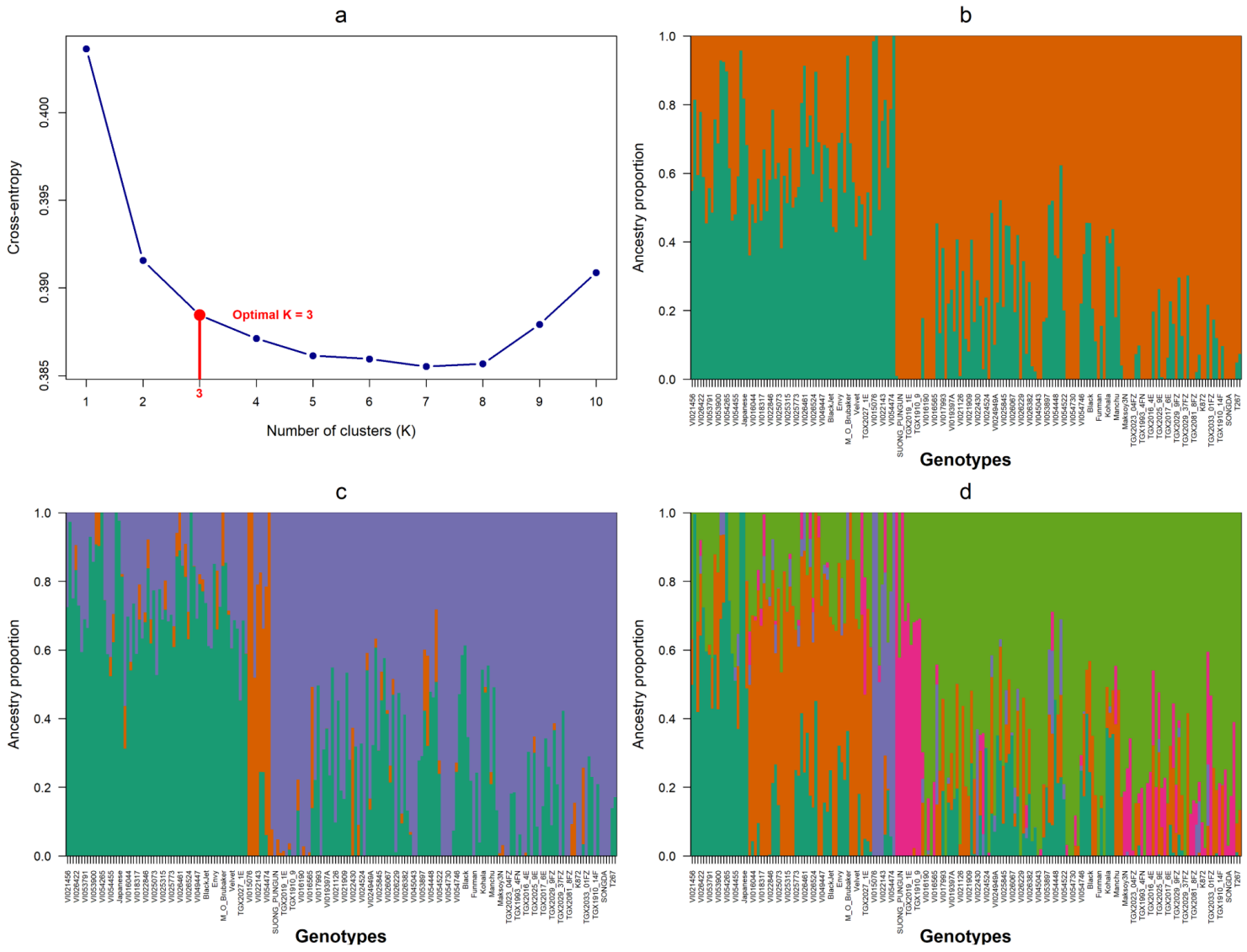

Figure 3

Population structure of the 191 soybean accessions inferred from admixture analysis. (a) Cross-entropy criterion used to evaluate the number of ancestral populations (K). (b-d) Estimated ancestry coefficients for individual accessions at K = 2 (b), K = 3 (c), and K = 5 (d). Each vertical bar represents one accession, and colors indicate the proportional membership of each accession to the inferred ancestral populations.

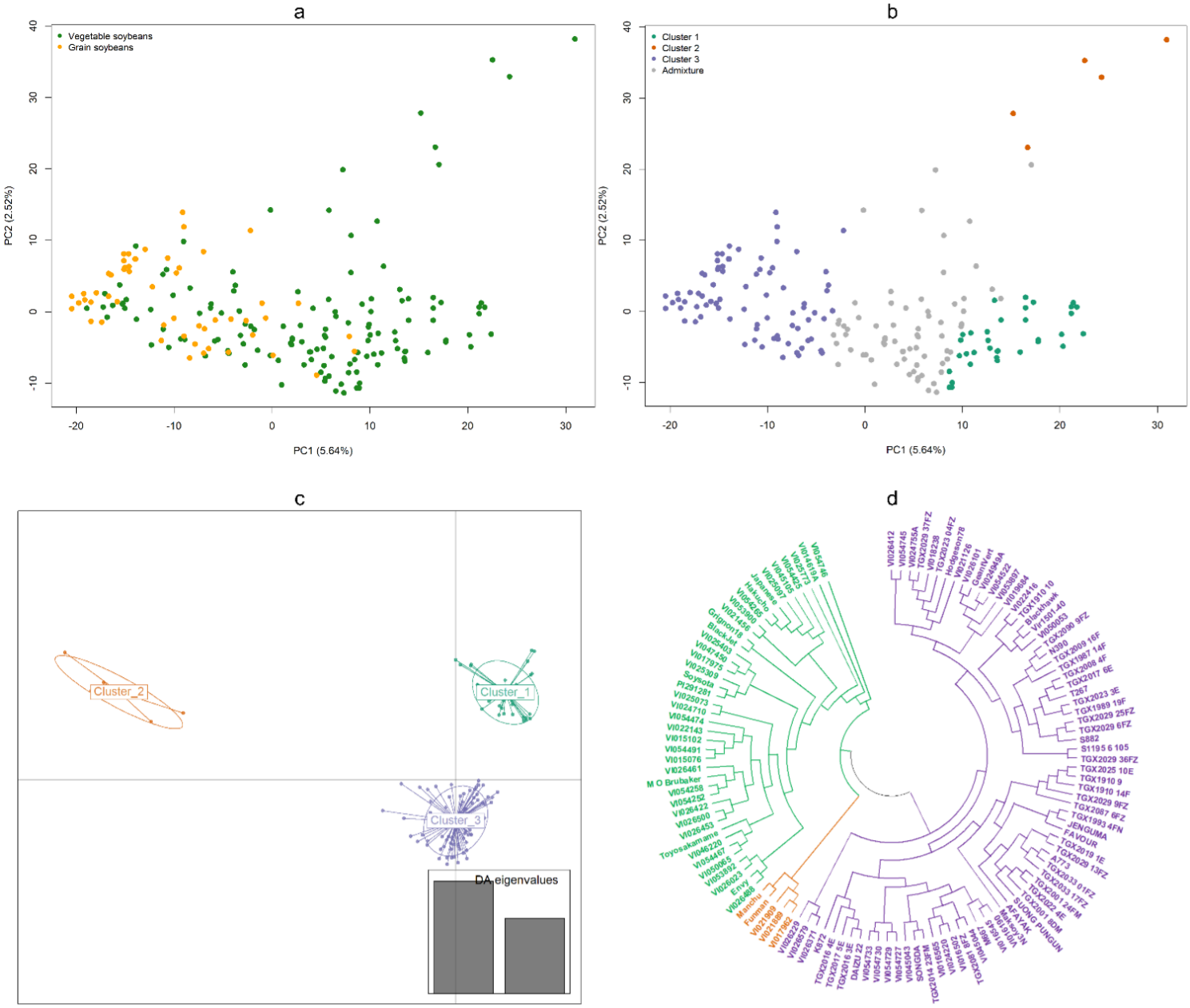

Figure 4

Genetic relationships among soybean accessions based on genome-wide SNP markers. (a) Principal components analysis of grain and vegetable soybean accessions based on genome-wide SNP markers; each point represents an individual accession and is colored according to soybean type (green: vegetable soybean; orange: grain soybean). (b) PCA showing the distribution of soybean accessions of the three genetic groups. (c) Population assignment inferred from DAPC. (d) Neighbor-joining phylogenetic tree illustrating the genetic relationships among the inferred genetic clusters. Colors represent the inferred genetic groups identified from the population structure analysis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.docx