Discovery and Genome-Wide Association of Single Nucleotide Polymorphic (SNP) Markers with Specific Biochemical Traits of Winged Bean (Psophocarpus tetragonolobus (L.) DC.)

Kunwar Sarvendra
National Botanical Research Institute
Umesh Chandra Sahoo
National Botanical Research Institute
Rayees Ahmad Lone
National Botanical Research Institute
Chandra Sekhar Mohanty
cs.mohanty.nbri@csir.res.in

National Botanical Research Institute

Research Article

Keywords:

Posted Date: June 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9326503/v1

Additional Declarations: No competing interests reported.

Abstract

Psophocarpus tetragonolobus (L.) DC is a nutritionally valuable legume. The present study aimed to characterize population structure, genetic diversity and marker–trait associations in 84 globally collected accessions of *P. tetragonolobus*. Single nucleotide polymorphic (SNPs) markers were identified using the reference-free UNEAK (Universal Network Enabled Analysis Kit) pipeline. The analysis resulted yielding 12,978 high-quality biallelic SNPs from over 80,000 tag pairs. Functional annotation via BLASTx revealed that approximately 16% of SNP-associated tags aligned with known proteins of *Glycine soja* and *Cajanus cajan*. Population structure analysis among the globally collected *P. tetragonolobus* using principal component analysis (PCA), STRUCTURE and phylogenetic reconstruction consistently resolved into six distinct genetic subpopulations. Biochemical evaluation of seed oil and protein content showed substantial variation across the accessions, the content ranged from (16.05–21.15) % and (15.70–22.89) %, respectively. Genome-wide association studies (GWAS) revealed 88 SNPs substantially linked with oil content of *P. tetragonolobus* and 54 with protein content. Of these, 14 and 16 SNPs were annotated to candidate genes involved in key regulatory and metabolic pathways. This study provides the first high-resolution genomic insight into *P. tetragonolobus* using genotyping-by-sequencing (GBS). The SNP markers and candidate genes identified serve as valuable resources for molecular breeding, enabling trait improvement and conservation of this underexploited legume.

Introduction

Winged bean (*Psophocarpus tetragonolobus* (L.) DC.) is an underutilized legume from the Fabaceae family (Bassal et al., 2020). It serves as a good source of protein (20.7–45.9) %, oil (7.2–21.5) % and carbohydrate (up to 35%). This makes it a valuable crop for both human consumption and animal feed (Kantha et al., 1984). This crop is tropical in distribution and thrives well in hot and humid regions of the world. It is a self-pollinating, diploid species (2n = 2x = 18) (Kharat et al., 2024).

Despite of its nutritional and ecological importance, *P. tetragonolobus* remains underutilized, due to limited genetic information (Dhaliwal et al., 2020). This limitation poses challenge for genome-level analysis and marker-assisted breeding. Previous studies assessed genetic diversity in *Psophocarpus tetragonolobus* using biochemical and molecular markers such as Inter Simple Sequence Repeat (ISSR) (Chen et al., 2015), Amplified Fragment Length Polymorphism (AFLP), and Random Amplified Polymorphic DNA (RAPD) (Mohanty et al., 2013). These studies offered preliminary insights into its genetic variability. However, these approaches lack a fine resolution and genome-wide coverages as provided by more advanced marker systems such as single nucleotide polymorphisms (SNPs). SNPs are highly abundant, biallelic markers and distributed throughout the genome (Tanzi, 2019).

Recent advancements in next-generation sequencing (NGS) technologies, particularly genotyping-by-sequencing (GBS), have enabled the generation of a large-scale SNP datasets for diverse crops, including those lacking reference genomes (Elshire et al., 2011). GBS has been successfully implemented in crops like: *Ziziphus jujube* (Chen et al., 2017), *Brassica rapa* (Bird et al., 2017) and

Triticum aestivum (Eltaher et al., 2018) and many more. GBS is employed not only for diversity studies but also for identifying loci associated with key agronomic traits. In oilseed and legume crops, GBS-based genome-wide association studies (GWAS) have covered critical loci for oil and protein content in *Glycine max* (Hwang et al., 2014; Bandillo et al., 2015; Jin et al., 2023), *Arachis hypogaea* (Umer et al., 2024), *Gossypium hirsutum* (Yuan et al., 2018), and *Brassica napus* (Liu et al., 2024). These studies, using high-density SNP panels and mixed linear models, have consistently highlighted the polygenic nature of seed composition traits and have revealed major Quantitative Trait Loci (QTLs) with potential breeding potential. The success of GBS-GWAS in identifying stable QTLs across environments and populations and deciphers its utility in underutilized crops.

In cases where reference genomes are incomplete or unavailable as is the case for *Phophocarpus tetragonolobus*—a non-reference-based pipeline such as UNEAK (Universal Network Enabled Analysis Kit) offers an alternative for SNP discovery. The UNEAK pipeline, implemented in TASSEL 3.0, identifies SNPs based on network filtering of tag pairs with single nucleotide differences, effectively by minimizing errors from repetitive or paralogous regions (Lu et al., 2013; Bradbury et al., 2007). UNEAK has been successfully employed in diploid and polyploid plants, including wheat and barley (Poland et al., 2012).

In the present study, the UNEAK pipeline was utilized to identify genome-wide SNPs from GBS data generated via Illumina sequencing in a globally diverse panel of 84 *P. tetragonolobus* accessions. The genetic diversity and population organization within the germplasm were then evaluated using these SNPs. In parallel, biochemical and phytochemical traits especially seed oil and protein content were evaluated among the studied accessions of *P. tetragonolobus.* GWAS were conducted to discover SNPs strongly linked with these traits. This provides an insight into the genetic loci controlling the protein and seed-oil composition in *P. tetragonolobus*.

This approach, aims to generate valuable genomic resources for *P. tetragonolobus* and to lay a foundation for future molecular breeding program targeting nutritional improvement in the underexploited legume *P. tetragonolobus*.

Materials and Methods

Plant materials

Seeds of 84 *Psophocarpus tetragonolobus* accessions were obtained from the ICAR– National Bureau of Plant Genetic Resources (NBPGR), India. Among these, 22 accessions were from India, 6 were from Ghana, 20 were from Papua New Guinea, 2 were from Indonesia, 1 were from the Philippines, and 33 were from Thailand. All genotypes were cultivated and maintained under uniform conditions in the experimental garden of CSIR– National Botanical Research Institute (NBRI), India. Detailed information for each accession, including country of origin and measured values for seed oil and protein content, is provided in **Supplementary Table S1**.

Extraction of oil

Approximately 250 g of dry seeds from each accession were crushed into fine powder using a domestic electric blender. Oil extraction was carried out using 100 g of seed powder and 500 mL of petroleum ether as solvent in a Soxhlet apparatus. This was operated continuously at 50°C for 4 hours. To extract the crude oil, the solvent was evaporated at reduced pressure using a rotary vacuum evaporator. The extracted oil was transferred to pre-weighed glass bottles, and the oil yield was calculated as a percentage of the initial dry seed weight. Before being analysed further, the samples were kept at 4°C.

Determination of protein

Protein was quantified using the Bradford technique (Bradford, 1976). 1 mL of 0.2M phosphate buffer solution (pH 7.2) was mixed with 0.05–0.1g of seed powder, and the mixture was then vortexed for 30 minutes. The resulting mixture was centrifuged for 10 minutes at 4°C at 5,000 rpm. The supernatant was collected, while the pellet was discarded. 50 µL of supernatant was mixed with 950 µL of Bradford reagent in the dark and the absorbance was recorded at 595 nm using a UV-Visible spectrophotometer (Thermo Evolution 260 Bio UV/VIS). The protein concentration was evaluated by comparing the absorbance to a standard curve produced from bovine serum albumin (BSA) at predetermined amounts.

DNA isolation, genotyping-by-sequencing and detection of SNPs

The Xcelgen Plant gDNA Isolation Kit (XG2611-01) was used to extract genomic DNA from seed samples, as per the manufacturer's protocol. The Qubit Fluorometer was used to measure the amount of DNA. Purity was confirmed by calculating the A260/280 ratio using a NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific).

Preparation of GBS libraries

Genotyping-by-sequencing (GBS) was carried out using the procedure reported by Elshire et al. (2011) with minor changes. To minimize genome complexity, 100 ng of high-quality genomic DNA from each sample was digested with the methylation-sensitive ApeKI restriction enzyme (recognition site: CGWGC, where W = A or T) at 75°C for 12 hours. T4 DNA ligase was used to ligate barcoded adapters (~ 0.1 pmol) and a common adapter to the sticky ends of the digested DNA fragments at 22°C for 180 minutes, and then heat inactivation was carried out. To create the final GBS library pools. The ligated products were combined into eight groups and amplified using PCR. The Agilent Bioanalyzer 2100 with a high sensitivity DNA chip was used to assess the quantity and quality of the libraries. Individual libraries were sequenced using paired-end chemistry (2 × 150 bp) on the Illumina HiSeq platform. The HiSeq 2500 machine was used to sequence each pooled library on a single lane. To ensure efficient adapter ligation in winged bean, the forward barcode adapter concentration was reduced to 0.06 pmol per 200 ng of

genomic DNA, as opposed to the 0.1 pmol commonly applied in wheat GBS workflows (Poland et al., 2012).

Analysis of GBS libraries using UNEAK pipeline

The UNEAK (Universal Network Enabled Analysis Kit) pipeline, included in the TASSEL 5.0 bioinformatic software, was used to analyse high-quality reads acquired after trimming adaptor sequences and deleting low-quality bases (Lu et al., 2013; Bradbury et al., 2007).

Identification and annotation of tag-pair and SNPs

SNPs were detected using the UNEAK pipeline, which is included in TASSEL version 5 (Lu et al., 2013). Following SNP identification, low-quality SNPs were filtered by removing loci with more than 60% missing data, while genotypes with more than 90% missing SNPs were eliminated from further analysis. High-quality SNPs retained after filtering. The data was used to generate HapMap format files using the UMapInfoToHapMapPlugin in TASSEL, and were subsequently converted into both biallelic and VCF (Variant Call Format) files for downstream applications (Dixon et al., 2014). To functionally annotate the tag pairs from which SNPs were obtained, NCBI's non-redundant (NR) protein database was searched using BLASTx with an E-value cutoff of 0.1. This requirement was used to ensure that the alignments met either a minimum of 60% identity across the entire tag length or 100% identity across. The filtered SNP genotyping data were then integrated with quantitative phytochemical trait data (protein content and seed oil) for GWAS study.

Population structure (Q matrix), kinship (K matrix) and principal component analysis (PCA) were incorporated into the GWAS model to account for potential confounding effects caused by population stratification and genetic relatedness among the global collected 84 *Psophocarpus tetragonolobus* accessions.

Population structure, and GWAS analysis

STRUCTURE v2.3.4 was used to calculate the population structure, with parallelization and automation enabled by Structure_threader (https://github.com/StuntsPT/Structure_threader). The ΔK approach, which was suggested by Evanno et al. (2005), was used to calculate the ideal number of clusters (K) when STRUCTURE was run under an admixture model as implemented in Structure Harvester (Earl and vonHoldt, 2012). This approach decomposes total genetic variance into and between-groups and within-group components to infer population stratification. The bar plot showing individual membership coefficients at the optimal K value (K = 6) was generated using Structure_threader. One of the six subpopulations was assigned to accessions with membership coefficients ≥ 0.70, whilst admixed accessions were those with lower coefficients. To complement the STRUCTURE analysis, with TASSEL v5.0, Principal Component Analysis (PCA) was carried out (Bradbury et al., 2007). The first three principal

components were extracted and visualized in a 3D scatter plot using the scatterplot3d package in R (Ligges et al., 2003). PCA results were merged with STRUCTURE-based subpopulation assignments to visualize clustering patterns. Phylogenetic analysis was performed using genome-wide SNP data formatted in HapMap and converted into compatible formats using TASSEL utilities. A pairwise genetic dissimilarity matrix was computed using PowerMarker v3.25 (Liu et al., 2005), a widely recognized software for population genetic analysis. The dissimilarity matrix was exported in PHYLIP format and used to construct a neighbor-joining (NJ) tree in PHYLIP v3.697 (Felsenstein, 1989). The resultant unrooted phylogenetic tree was visualized using FigTree v1.4.3. Kinship matrix estimation was conducted in TASSEL v5.0, based on SNP genotype data. Kinship was determined as the proportion of shared alleles among individuals and used as a measure for relatedness in future genome-wide association studies (Endelman etal., 2012). A heatmap representing the kinship distance matrix was generated using the pheatmap package in R (Kolde, 2019).

Marker trait association

TASSEL v5.0 (Evolution, Trait Analysis by Association and Linkage), a statistical software intended for association mapping in structured populations, was used to perform genome-wide association analysis (GWAS). Individual SNP markers and the measured phytochemicals i.e, protein content and seed oil were correlated using the Mixed Linear Model (MLM) and General Linear Models (GLM). To adjust the false positives, the MLM considered the relatedness among genotypes (K matrix) and population structure (Q matrix). The statistical robustness of marker-trait associations was assessed by comparing observed and expected $-\log_{10}(P)$ values using quantile–quantile (Q–Q) plots. According to the Benjamini–Hochberg method, multiple testing was considered with a false discovery rate (FDR) threshold of ≤ 0.05 (Benjamini and Hochberg, 1995). SNPs surpassing the FDR-adjusted p-value threshold were considered significantly correlated with the desired trait.

Blast search for Putative Candidate Genes

To determine the potential candidate genes associated with important SNP markers, BLASTn searches were conducted using the flanking sequences of associated SNPs as queries. Searches were performed against the *Glycine max* genome (version Mt4.0 v17) using public databases, including NCBI and Phytozome. Candidate genes were inferred based on sequence similarity and functional annotation. Genes located near the significant SNPs and involved in relevant metabolic or regulatory pathways were prioritized as contributors to the variation in protein and seed oil content in *P. tetragonolobus*.

Results

Variation in Protein and Oil Content

Seed oil and protein content were quantified among the 84 globally collected *P. tetragonolobus* accessions sourced from six countries. Oil content ranged from 16.05% to 21.15%, with IC 95238 (India) showing the highest oil concentration. Protein content varied from 15.70% to 22.89%, with EC 11885

(Ghana) showing the highest value. These results highlighted substantial biochemical diversity in phytochemical traits among the accessions **(Supplementary Table S1)**. This forms a strong basis for identifying trait-linked genetic markers in subsequent analyses.

Phytochemical Trait Distribution

Analysis of oil and protein content across 84 *Psophocarpus tetragonolobus* accessions revealed significant quantitative variation, with discernible patterns associated with geographic origin. Notably, Indian accessions exhibited a wide range of protein levels, with some lines such as IC 95241 (21.80%) and IC 95238 (20.27%) among the highest protein-containing entries. Similarly, Ghanaian lines like EC 11885 and EC 27885-2 showed elevated protein values (22.55% and 20.39%, respectively), highlighting their potential as nutritionally superior germplasm. Papua New Guinean accessions were particularly enriched for seed oil content. For instance, IC 95238 from India showed the highest oil content recorded (21.15%), while multiple Papua New Guinea accessions, including EC 38956-1 (18.71%), EC 38948 (18.57%), and EC 38824-2 (18.35%), consistently exhibited oil contents above 18%. The lowest protein content was observed in EC 38957 (16.18%) from Papua New Guinea, and the lowest oil content (16.05%) in EC 178282 from Thailand. This geographic clustering of extreme phenotypes provides a valuable foundation for dissecting the genetic architecture of seed phytochemical traits via SNP-based genome-wide association studies (GWAS). Full trait details and origins for all accessions are provided in **Supplementary Table S1**.

Genomic DNA Quality and Processing

High-quality genomic DNA was successfully extracted from seed material. Agarose gel electrophoresis was used to verify the integrity of the DNA, and Qubit Fluorometry and Nanodrop spectrophotometry were used to ascertain its concentration and purity. The quality was appropriate for genotyping-by-sequencing (GBS) and downstream library preparation. For complexity reduction and efficient library construction, DNA samples were individually barcoded and subsequently pooled into eight multiplexed libraries prior to sequencing.

GBS Data Generation

The methylation-sensitive *ApeKI* restriction enzyme was used to create genotyping-by-sequencing (GBS) libraries, which were individually barcoded and subsequently pooled to reduce complexity and improve sequencing efficiency. Following adapter ligation, the pooled libraries were subjected to PCR amplification and quality control using the Agilent Bioanalyzer to ensure proper fragment size distribution and concentration. High-throughput sequencing was then performed on the Illumina HiSeq 2500 platform using 2 × 150 bp paired-end chemistry, yielding high-resolution data per accession.

A total of 84 *P. tetragonolobus* accessions were sequenced, yielding between 170,782 and 17,277,548 paired-end reads per accession. The total number of raw reads across all samples exceeded 670 million, and the combined dataset spanned approximately 102.3 gigabases (Gb), with individual accession-level outputs ranging from 26 Mb to 4.75 Gb. The average sequencing depth per accession was approximately

1.2 Gb, though variation in library size and read count reflected the biological and technical diversity across samples.

Sequencing statistics, including total reads, base pairs, and data volume per accession, are detailed in Supplementary Table S2. These high-quality sequencing reads served as input for SNP discovery using the reference-free UNEAK pipeline, making the dataset suitable for downstream population structure analysis, diversity assessment and GWAS analysis.

SNP Discovery via UNEAK Pipeline

SNP identification was carried out using the reference-independent UNEAK pipeline in TASSEL v5.0. A total of 167,845 tag pairs, representing 335,690 unique sequence tags, were generated from the raw sequencing reads and used for downstream variant discovery. After applying network-based filtering to remove low-confidence and redundant tags, 80,941 tag pairs were retained. Corresponding to 161,882 unique tags containing putative SNPs. Only SNPs with a minor allele frequency (MAF) between 0.05 and 0.5 were considered for further analysis. The final SNP dataset consisted of 12,978 high-quality, biallelic SNPs, which were exported in HapMap, biallelic HapMap, and VCF formats for further use in population structure analysis, phylogenetic tree construction, and GWAS study. A schematic representation of the UNEAK pipeline workflow is provided in Fig. 1. Summary statistics for the filtered SNPs, including allele frequency range, call rate thresholds, and heterozygosity, are presented in Table 3. SNPs with more than 60% missing data were removed, and a minimum call rate of 0.6 was applied to ensure data reliability.

Table 3

Functional annotation of significant SNPs associated with oil and protein content in *Psophocarpus tetragonolobus* based on GWAS analysis using the UNEAK pipeline. SNPs were annotated via BLASTx search against the NCBI NR database. The table includes marker ID, alleles, F-values, p-values, marker R^2, and associated protein/gene functions.

Trait	Significant annotated SNPs Marker	Allele	F	p-value	Marker R^2	Sequence discription
Oil	TP7454	C/T	11.34126	9.57E-05	0.28937	Pumilio-like 2 isoform C
	TP167215	C/T	10.72561	1.37E-04	0.28595	protein OSB1, mitochondrial-like
	TP7699	A/G	9.56845	2.04E-04	0.19745	separase-like
	TP36378	G/T	9.47727	2.19E-04	0.19397	probable galacturonosyltransferase 10
	TP136011	C/T	9.80856	2.61E-04	0.25872	Dof zinc finger partial
	TP165525	C/T	9.21883	2.79E-04	0.19859	kinesin-4
	TP160758	A/G	9.00196	3.24E-04	0.18799	exocyst complex component EXO70A1-like
	TP74263	C/T	8.71577	4.12E-04	0.18811	inositol 3-alpha-galactosyltransferase
	TP154831	A/G	8.81166	4.88E-04	0.2364	probable amino acid permease 7
	TP12771	A/T	8.39147	5.24E-04	0.17764	glycerol-3-phosphate transporter partial
	TP137226	C/T	8.36414	5.79E-04	0.19338	uncharacterized protein LOC106797271
	TP125930	C/G	8.40601	6.87E-04	0.22207	dnaJ homolog subfamily C member 17
	TP143097	C/T	8.0514	8.84E-04	0.2199	unnamed protein product, partial
	TP109600	C/T	7.68398	9.45E-04	0.16611	PREDICTED: uncharacterized protein LOC108840324 isoform X2
Protein	TP32891	A/G	12.3117	4.66E-05	0.26615	thaumatin 1
	TP22600	A/G	10.51155	1.05E-04	0.22314	uncharacterized protein LOC100499880

Trait	Significant annotated SNPs Marker	Allele	F	p-value	Marker R^2	Sequence discription
	TP81129	C/T	10.11787	1.47E-04	0.20601	hypothetical protein GLYMA_03G218600
	TP148163	C/T	9.59765	1.99E-04	0.19549	CASP 1D2
	TP154752	A/C	9.9622	2.03E-04	0.22296	F-box At5g07670-like
	TP152560	C/T	14.05563	3.56E-04	0.15482	signal recognition particle receptor subunit alpha-like
	TP7832	A/T	8.86733	4.76E-04	0.21641	Glutamate dehydrogenase 2 isoform F
	TP40987	C/T	8.48655	4.91E-04	0.17563	chitinase
	TP39131	C/T	8.37452	5.48E-04	0.16824	Ubiquitin carboxyl-terminal hydrolase 26
	TP163405	C/T	8.53526	6.12E-04	0.20569	pentatricopeptide repeat-containing chloroplastic
	TP28750	C/T	8.04609	6.95E-04	0.16482	ABC transporter C family member 10-like isoform X2
	TP119027	A/G	13.10425	7.08E-04	0.16873	Folate transporter chloroplastic
	TP51308	A/T	7.99614	7.11E-04	0.16633	hypothetical protein CR513_03284, partial
	TP161016	C/T	7.74198	9.01E-04	0.16559	hypothetical protein PHAVU_001G2478001g
	TP137159	C/G	7.84455	9.53E-04	0.18867	Multidrug and toxin extrusion 1

Functional Annotation of SNP-Linked Tags

To learn more about the potential biological functions of the identified SNP loci, tag pairs were BLASTx analysed against the NCBI non-redundant (NR) protein database with an E-value cutoff of 0.1. This criterion was selected to ensure high-confidence alignments by retaining sequences with at least 60% identity across the whole tag length or 100% identity over at least half of the sequence. From the 161,882 SNP-containing tag sequences generated through the UNEAK pipeline, a total of 25,634 tag pairs (~ 16%) were successfully annotated with known protein functions. These results are summarized in Table 1. Taxonomic distribution of the BLAST hits revealed that most top matches belonged to *Glycine soja*, followed by *Cajanus cajan* and *Glycine max*. The analyzed data is consistent with the phylogenetic

proximity of these species to *P. tetragonolobus* (Fig. 2a). Additionally, the total number of tag pairs, the number of successfully annotated sequences, and the overall annotation rate are presented in Fig. 2b. These annotated SNP-linked tags form a valuable resource for identifying candidate genes associated with important phytochemical traits.

Table 1
Summary statistics of filtered SNP variants generated by the UNEAK pipeline in *Psophocarpus tetragonolobus*.

Sr no.	Pipeline	Number of variants	Site depth	Site missing	Mean MAF	Heterozygosity	% heterozygosity ≥ 0.5
1.	UNEAK	12978	41.8×	10.1%	0.05	0.32	26.83%

Population Structure and Principal Component Analysis

The genetic structure of the 84 *P. tetragonolobus* accessions was examined by using 18,768 high-quality SNP markers. The number of genetic clusters (K) was estimated using STRUCTURE, and the results ranged from 1 to 10. Using STRUCTURE Harvester, the ΔK technique of Evanno et al. (2005) determined that K = 6 and was reported to be the ideal number of clusters. The presence of six different subpopulations was suggested by a strong peak at K = 6 in the ΔK distribution plot (Fig. 3a).

To validate this structure, PCA was performed. The results were visualized in a 3D PCA scatter plot (Fig. 3b). A clear clustering of the 84 accessions into six groups was observed. These groups corresponded closely with the subpopulations inferred by STRUCTURE, highlighting the underlying genetic differentiation and population stratification.

Individual ancestry proportions were visualized through a STRUCTURE bar plot at K = 6 (Fig. 3c). Each vertical bar represented one accession, subdivided into color-coded segments that reflected the proportion of its genome assigned to each subpopulation. Accessions with membership probabilities ≥ 0.70 were classified into specific subpopulations, while those with lower values were considered admixed. The complete membership coefficient matrix and final subgroup assignments were provided in Supplementary Table 3.

Phylogenetic Analysis of Accessions

For further validation of genetic differentiation and population structure, a phylogenetic tree was constructed using the genome-wide SNP data. A pairwise genetic distance matrix was calculated using PowerMarker based on the filtered SNP dataset and a neighbor-joining tree was generated using PHYLIP. The resulting tree was visualized in FigTree v1.4.3. The unrooted phylogenetic tree (Fig. 4) revealed well-defined clades that corresponded to the six subpopulations previously identified by STRUCTURE analysis. Accessions with high membership probabilities in a single cluster were grouped together, whereas admixed individuals were positioned on intermediate branches, reflecting shared ancestry. Strong consistency among the STRUCTURE analysis, PCA clustering, and phylogenetic relationships was

observed, supporting the presence of robust genetic stratification within the *P. tetragonolobus* diversity panel.

GWAS for Protein and Oil Traits and Kinship Matrix

To uncover the genetic loci associated with protein and seed oil content, GWAS was conducted using the Mixed Linear Model (MLM) in TASSEL 5, which incorporated both kinship (K-matrix) and population structure (Q-matrix) to control for confounding due to relatedness and stratification. 88 SNPs in all were discovered to be substantially correlated with oil content, and 54 SNPs with protein content, based on a threshold of $-\log_{10}(p) \geq 3.0$. A summary of the total and annotated SNPs for each trait is presented in Table 2. Out of these, 14 SNPs for oil and 16 SNPs for protein traits were successfully annotated to known genes using BLASTx. The kinship matrix, calculated from genome-wide SNPs, revealed varying degrees of genetic similarity among the accessions and was visualized as a heatmap in Fig. 5. This matrix was used as a random effect in the MLM to account for familial relatedness. To evaluate the statistical performance of the model and detect potential inflation, quantile–quantile (Q–Q) plots were generated for both traits. The observed versus expected $-\log_{10}$(p-values) plots showed minimal deviation from the diagonal line, with clear outliers in the upper tail indicating true SNP-trait associations. These plots are shown in Fig. 6a (oil) and Fig. 6b (protein).

Table 2
Total number of significant and annotated SNPs associated with oil and protein traits at p-value < 0.001.

Traits	# Significant SNPs	# Significant annotated SNPs
Oil	88	14
Protein	54	16

Identification of Putative Candidate Genes from Significant SNPs

The important SNPs found by GWAS were functionally annotated using BLASTx against the NCBI non-redundant protein database in order to identify biologically significant factors linked to protein and oil content traits. A total of 14 oil-associated and 16 protein-associated SNPs were successfully annotated (Table 3). These SNPs were linked to candidate genes involved in regulatory and metabolic processes, including Pumilio-like RNA-binding proteins, mitochondrial OSB1-like proteins, galacturonosyltransferases, and cytochrome c biogenesis proteins, among others. These gene functions suggest a plausible molecular basis for variation in oil and protein biosynthesis pathways in *P tetragonolobus*. These candidate loci offer promising targets for future validation studies and marker-assisted breeding programs aimed at improving seed quality traits in underutilized legume *P tetragonolobus*.

Discussion

This study demonstrates the effectiveness of genotyping-by-sequencing (GBS) for exploring the genetic architecture of phytochemical traits in the underutilized legume *Psophocarpus tetragonolobus*. Utilizing the UNEAK pipeline within TASSEL v5.0 enabled to discover 80,941 high-quality SNPs from 84 accessions without relying on a reference genome. UNEAK's reference-independent framework—where raw reads are collapsed into 64 bp tags and then compared for single-nucleotide differences—proves especially valuable in species like *P. tetragonolobus* where whole-genome assemblies are unavailable or incomplete (Lu et al., 2013).

UNEAK's stringency is achieved through its network-based filtering algorithm, which eliminates redundant and ambiguous tags, ensuring confidence in SNP calling (Lu et al., 2013). Similar approaches have been successfully implemented in cowpea (Xiong et al., 2016), groundnut (Umer et al., 2024), and soybean (Bandillo et al., 2015), reinforcing its utility for genomic studies in non-model legumes. In switchgrass, UNEAK also yielded high-quality markers, validating its applicability across diverse species with complex genomes (Lu et al., 2013).

The methylation-sensitive restriction enzyme *ApeKI*, known for targeting low-copy genomic regions and avoiding repetitive DNA, was employed during GBS library preparation to ensure coverage of informative loci (Poland et al., 2012). This strategy enhances sequencing efficiency and genome representation, as shown in other legume crops (Boutet et al., 2016; Hwang et al., 2014). To improve data quality, filtering thresholds such as a minor allele frequency (MAF) of 0.05 and call rate of ≥ 0.6 were applied, resulting in a final dataset of 12,978 robust, biallelic SNPs. These parameters are in line with established best practices for reliable genome-wide association studies (Yu et al., 2006).

The SNP dataset revealed a transition-to-transversion bias, with transitions occurring more frequently—a phenomenon consistent with earlier reports in soybean (Jin et al., 2023), *Brassica napus* (Kim et al., 2024), and cotton (Yuan et al., 2018). This mutational trend likely reflects evolutionary tolerance for transitions over transversions, further validating the biological integrity of our variant dataset.

To address confounding factors such as population stratification, multiple complementary methods were applied: STRUCTURE analysis, Principal Component Analysis (PCA), and phylogenetic reconstruction. STRUCTURE analysis (Evanno et al., 2005; Earl & vonHoldt, 2012) using the ΔK method identified six genetically distinct subpopulations (K = 6), a pattern confirmed by PCA and a neighbor-joining phylogenetic tree derived from SNP-based pairwise dissimilarity matrices computed using PowerMarker. This convergence among independent methods strongly supports robust genetic differentiation within the global *P. tetragonolobus* panel—mirroring findings in soybean diversity studies (Bandillo et al., 2015; Jin et al., 2023).

The Mixed Linear Model (MLM) framework, incorporating both kinship (K-matrix) and population structure (Q-matrix), was used to enhance GWAS accuracy and reduce false-positive associations (Yu et al., 2006). This statistical rigor is critical when dissecting complex traits like oil and protein content, and

similar MLM-based GWAS models have been effectively used in groundnut (Umer et al., 2024), soybean (Jin et al., 2023), and rapeseed (Kim et al., 2024).

Despite potential limitations related to sequencing depth or adapter quality, the pipeline produced evenly distributed SNPs across samples, with high confidence and low missingness. This highlights the robustness of UNEAK when applied to underutilized legumes and reaffirms its capacity to generate biologically meaningful results even in the absence of a reference genome (Lu et al., 2013; Poland et al., 2012).

Ultimately, this study not only identified high-confidence SNP markers associated with oil and protein content but also uncovered candidate genes potentially involved in their regulation. These include genes encoding Pumilio-like RNA-binding proteins, mitochondrial OSB1-like proteins, and cytochrome c biogenesis factors, which have been implicated in seed composition and metabolic regulation in other legume systems (Hwang et al., 2014; Bandillo et al., 2015). This genomic resource lays the foundation for future functional validation and offers significant potential for marker-assisted breeding in *P. tetragonolobus*.

Conclusion

This study presents a comprehensive genome-wide assessment of genetic diversity, population structure, and trait-associated markers in *Psophocarpus tetragonolobus* using genotyping-by-sequencing (GBS). Using the UNEAK pipeline in TASSEL v5.0, over 12,000 high-quality SNPs were identified across 84 accessions, which enabled the discovery of 88 SNPs associated with oil content and 54 with protein content. Several significant SNPs were linked to functionally relevant genes such as Pumilio-like RNA-binding proteins and cytochrome c biogenesis factors. STRUCTURE analysis, principal component analysis, and phylogenetic reconstruction consistently revealed six genetically distinct subpopulations, highlighting strong population stratification. A Mixed Linear Model (MLM) incorporating both kinship and population structure was applied to reduce false positives in genome-wide association analysis. Overall, the study provides valuable genomic resources and trait-linked markers that can accelerate marker-assisted selection and trait improvement in *P. tetragonolobus*, contributing to its broader genetic enhancement and utilization.

Abbreviations

SNP
Single Nucleotide Polymorphism
GBS
Genotyping–by–Sequencing
GWAS
Genome–Wide Association Study
PCA

Principal Component Analysis
QTL
Quantitative Trait Locus
BLASTx
Basic Local Alignment Search Tool for translated proteins
NR
Non–Redundant (NCBI database)
MAF
Minor Allele Frequency
R^2
Coefficient of Determination

Declarations

Conflict of Interest

The authors declare that they have no known financial or non-financial conflicts of interest that could have appeared to influence the work reported in this paper. All co-authors have reviewed and approved the final version of the manuscript.

Ethical Approval

Not applicable.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Funding

The work was supported by the in-house grant of CSIR- National Botanical Research Institute.

Author Contribution

KS designed the study, performed the SNP discovery and GWAS analysis, and prepared the initial draft of the manuscript. UCS conducted biochemical assays, contributed to data interpretation, and assisted in result validation. RAL contributed to STRUCTURE and phylogenetic analysis, annotation of candidate genes, and visualization of results. CSM conceptualized and supervised the project, provided critical

revisions, and approved the final manuscript for submission. All authors read and approved the final manuscript.

Acknowledgments

The authors acknowledge Director, CSIR- National Botanical Research Institute for providing facilities to carry out the research. The manuscript number provided by the institute is CSIR-NBRI_MS/2025/07/12.

Data Availability

Raw sequencing data are available in the NCBI Sequence Read Archive (SRA) under Bio Project accession number **PRJNA1287067** .

References

1. Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A (2015) A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8(3):plantgenome2015–plantgenome2004
2. Bassal H, Merah O, Ali AM, Hijazi A, El Omar F (2020) Psophocarpus tetragonolobus: An underused species with multiple potential uses. Plants 9(12):1730
3. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol 57:289–300
4. Bird KA, An H, Gazave E, Gore MA, Pires JC, Robertson LD, Labate JA (2017) Population structure and phylogenetic relationships in a diverse panel of Brassica rapa L. Front Plant Sci 8:321
5. Boutet G, Alves Carvalho S, Falque M, Peterlongo P, Lhuillier E, Bouchez O, Lavaud C, Pilet-Nayel ML, Rivière N, Baranger A (2016) SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics 17:1–14
6. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635
7. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72(1–2):248–254
8. Chen D, Yi X, Yang H, Zhou H, Yu Y, Tian Y, Lu X (2015) Genetic diversity evaluation of winged bean (Psophocarpus tetragonolobus (L) DC.) using inter-simple sequence repeat (ISSR). Genet Resour Crop Evol 62:823–828
9. Chen W, Hou L, Zhang Z, Pang X, Li Y (2017) Genetic diversity population structure and linkage disequilibrium of a core collection of Ziziphus jujuba assessed with genome-wide SNPs developed by genotyping by-sequencing and SSR markers. Front Plant Sci 8:575

10. Dhaliwal SK, Talukdar A, Gautam A, Sharma P, Sharma V, Kaushik P (2020) Developments and prospects in imperative underexploited vegetable legumes breeding: a review. Int J Mol Sci 21(24):9615
11. Dixon A, Comont D, Slavov GT, Neve P (2021) Population genomics of selectively neutral genetic structure and herbicide resistance in UK populations of Alopecurus myosuroides. Pest Manag Sci 77(3):1520–1529
12. Earl DA, Von Holdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evano method. Conserv Genet Resour 4:359–366
13. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5):e19379
14. Eltaher S, Sallam A, Belamkar V, Emara HA, Nower AA, Salem KF, Poland J, Baenziger PS (2018) Genetic diversity and population structure of F3:6 Nebraska winter wheat genotypes using genotyping-by-sequencing. Front Genet 9:76
15. Endelman JB, Jannink JL (2012) Shrinkage estimation of the realized relationship matrix. G3. Genes Genomes Genet 2(11):1405–1413
16. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14(8):2611–2620
17. Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). Cladistics 5:164–166
18. Hwang EY, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15(1):1
19. Jin H, Yang X, Zhao H, Song X, Tsvetkov YD, Wu Y, Gao Q, Zhang R, Zhang J (2023) Genetic analysis of protein content and oil content in soybean by genome-wide association study. Front Plant Sci 14:1182771
20. Kantha SS, Erdman JW (1984) The winged bean as an oil and protein source: A review. J Am Oil Chem Soc 61(3):515–525
21. Kharat MA, Bhise DR, Galande HR (2024) WINGED BEAN. Beans: An insight study 201
22. Kim WJ, Yang B, Kim DG, Kim SH, Lee YJ, Kim J, Baek SH, Kang SY, Ahn JW, Choi YJ, Bae CH (2024) Genotyping-by-sequencing analysis reveals associations between agronomic and oil traits in gamma ray-derived mutant rapeseed (Brassica napus L). Plants 13(11):1576
23. Kolde R (2019) Pheatmap: pretty heatmaps. R package version 1(2):726
24. Ligges U, Mächler M (2003) Scatterplot3d-an R package for visualizing multivariate data. J Stat Softw 8:1–20
25. Liu H, Yuan Y, Tang Y, Li R, Ye K, Zhang M, Lu K, Yin N, Zhao H, Liu Y, Huang T (2024) Genome- and transcriptome-wide association studies reveal the genetic basis of seed palmitic acid content in Brassica napus. Journal of Integrative Agriculture
26. Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9):2128–2129

27. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH et al (2013) Switchgrass genomic diversity ploidy and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet 9:e1003215
28. Mohanty CS, Verma S, Singh V, Khan S, Gaur P, Gupta P, Nizar MA, Dikshit N, Pattanayak R, Shukla A, Niranjan A (2013) Characterization of winged bean (Psophocarpus tetragonolobus (L) DC.) based on molecular chemical and physiological parameters
29. Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5(3)
30. Tanzi AS (2019) Genetic investigation of plant architecture yield and diversity in winged bean (Psophocarpus tetragonolobus (L) DC.) for breeding programmes (Doctoral dissertation University of Nottingham)
31. Umer MJ, Lu Q, Huang L, Batool R, Liu H, Li H, Wang R, Qianxia Y, Varshney RK, Pandey MK, Hong Y (2024) Genome-wide association study reveals the genetic basis of amino acids contents variations in Peanut (Arachis hypogaea L). Physiol Plant 176(5):e14542
32. Xiong H, Shi A, Mou B, Qin J, Motes D, Lu W, Ma J, Weng Y, Yang W, Wu D (2016) Genetic diversity and population structure of cowpea (Vigna unguiculata L. Walp). PLoS ONE 11(8):e0160941
33. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203–208
34. Yuan Y, Wang X, Wang L, Xing H, Wang Q, Saeed M, Tao J, Feng W, Zhang G, Song XL, Sun XZ (2018) Genome-wide association study identifies candidate genes related to seed oil composition and protein content in Gossypium hirsutum L. Front Plant Sci 9:1359

Figures

Figures and legends

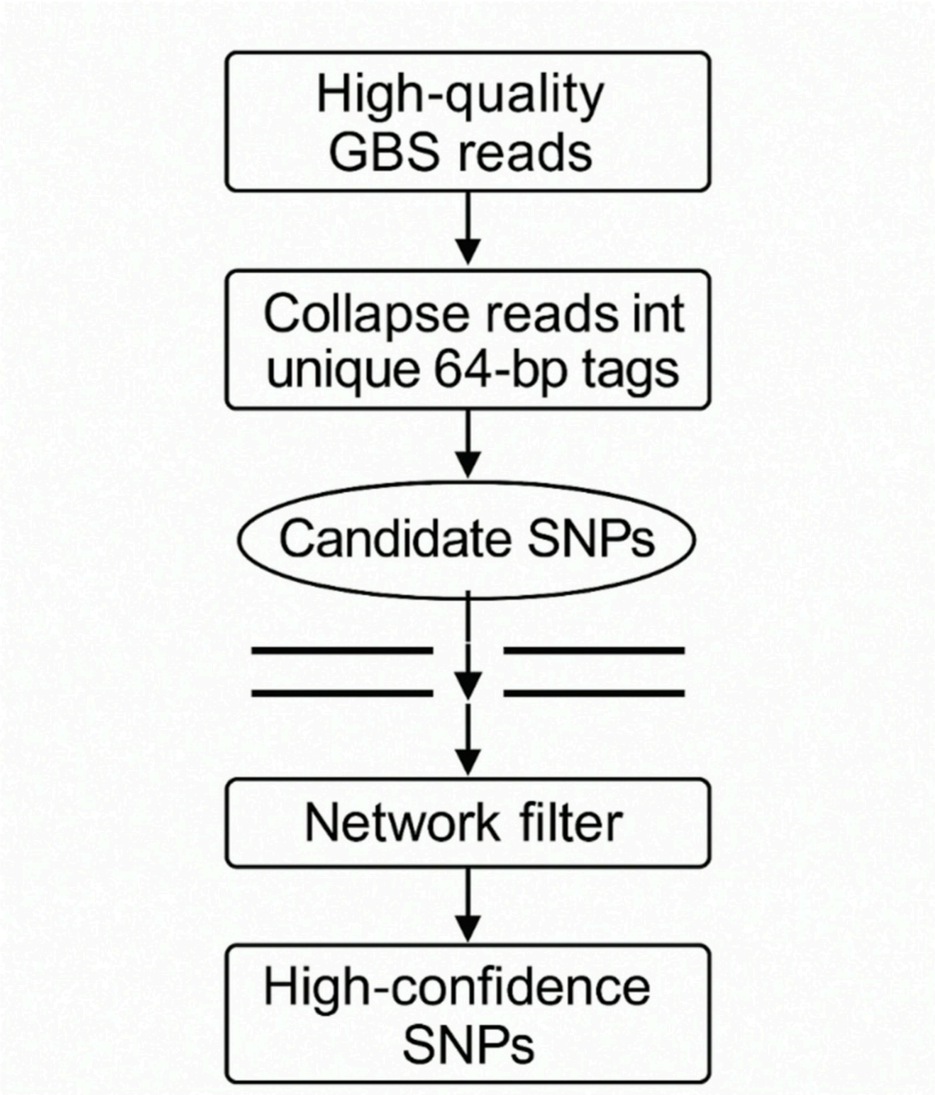

Figure 1

Schematic overview of the UNEAK pipeline used for SNP discovery in *Psophocarpus tetragonolobus*. The pipeline, implemented in TASSEL 5.0, processes high-quality GBS reads by collapsing them into unique 64-bp tags. Tags are aligned pairwise to identify single nucleotide mismatches (candidate SNPs), which are filtered via a network-based algorithm to remove sequencing errors, paralogs, and repetitive sequences. The resulting high-confidence SNPs are exported in HapMap and VCF formats for downstream population structure and GWAS analysis.

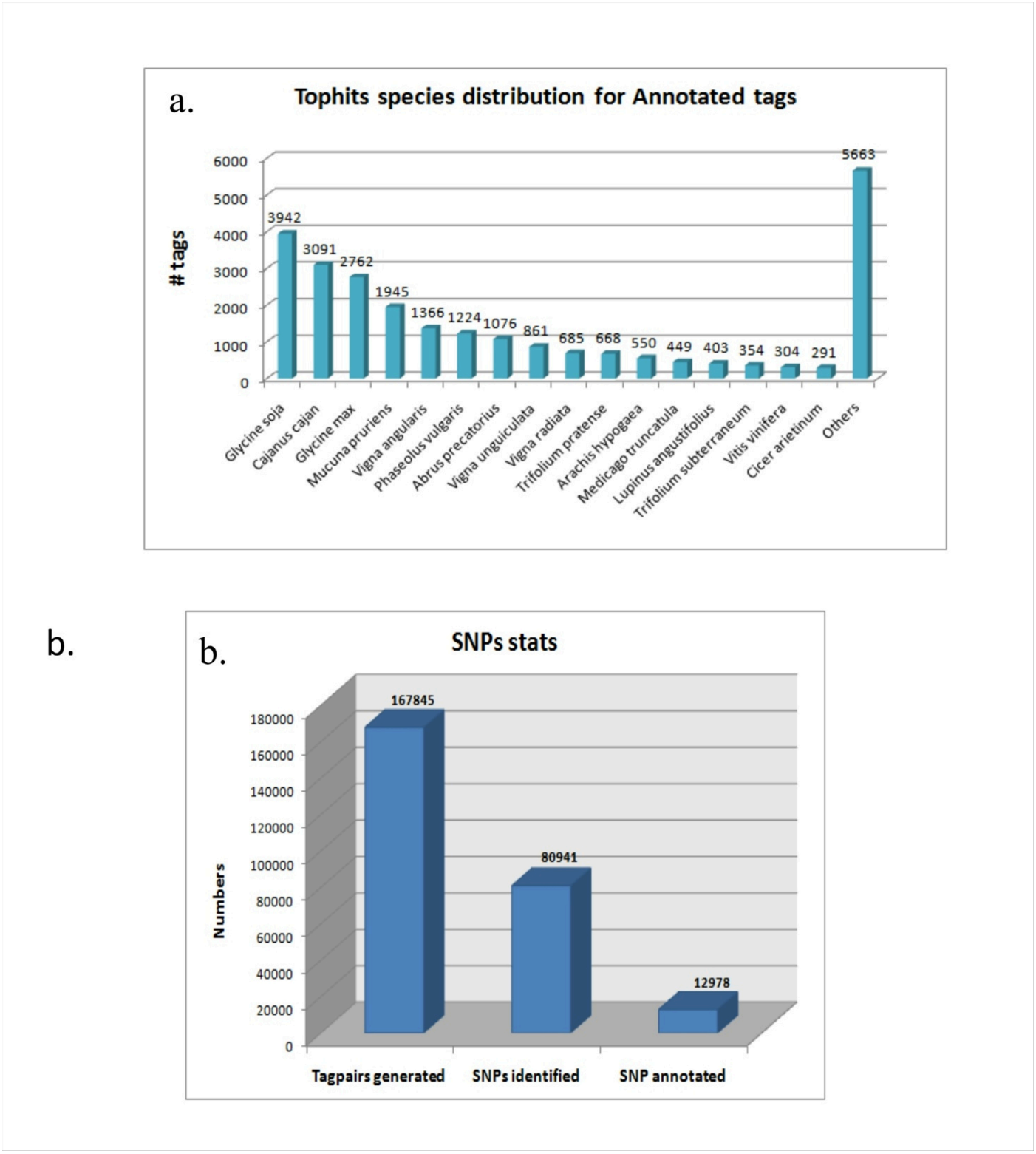

Figure 2

Annotation of tag pairs derived from UNEAK pipeline SNP discovery.**2(a)** Distribution of top-hit species from BLASTx annotation of tag sequences. Tag pairs containing SNPs were aligned to the NCBI non-redundant (NR) protein database with an E-value cutoff of 0.1. The majority of significant hits were to proteins from *Glycine soja*, followed by *Cajanus cajan* and *Glycine max*, reflecting the evolutionary proximity of these legumes to *Psophocarpus tetragonolobus*. **2(b)** Summary of tag pair annotation

results. The bar chart displays the total number of tag pairs, the number that yielded significant BLASTx hits (annotated), and the percentage of total tag pairs successfully annotated.

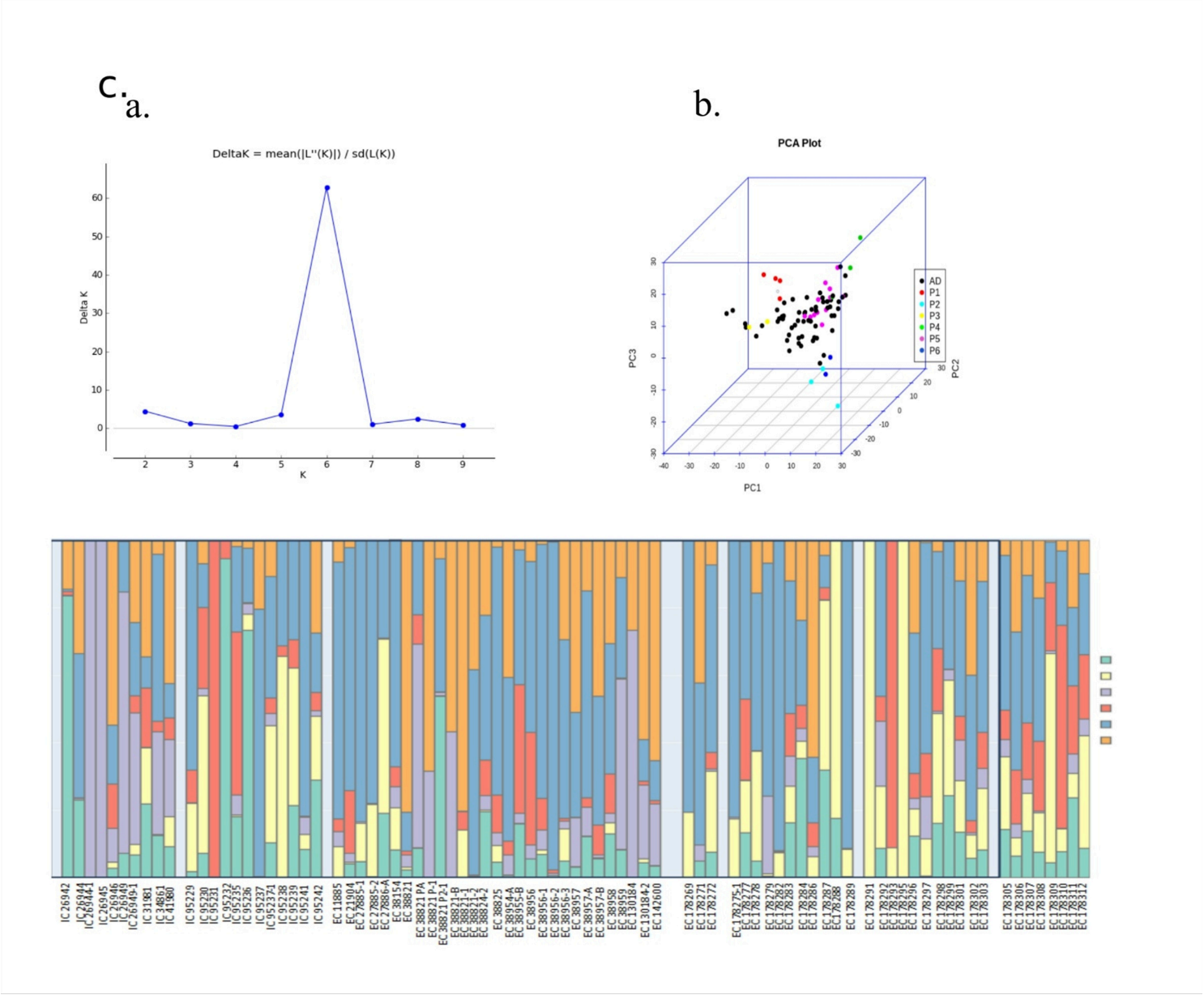

Figure 3

Population structure and genetic differentiation among 84 *Psophocarpus tetragonolobus*accessions. (a) ΔK plot for determining the optimal number of subpopulations. The peak at K = 6 was identified using the Evanno method, indicating the most likely number of genetic clusters in the panel. (b) Principal Component Analysis (PCA) of genotypic data based on 18,768 SNPs. The 3D scatter plot displays genetic relationships among accessions, with each point representing an individual grouped by population cluster. The clustering pattern supports the structure results. (c) STRUCTURE bar plot at K = 6 showing individual ancestry proportions. Each vertical bar represents an accession, partitioned into colored segments reflecting the proportion of its genome derived from each of the six inferred

subpopulations. Accessions with membership probability ≥ 0.70 were assigned to a specific cluster; others were considered admixed.

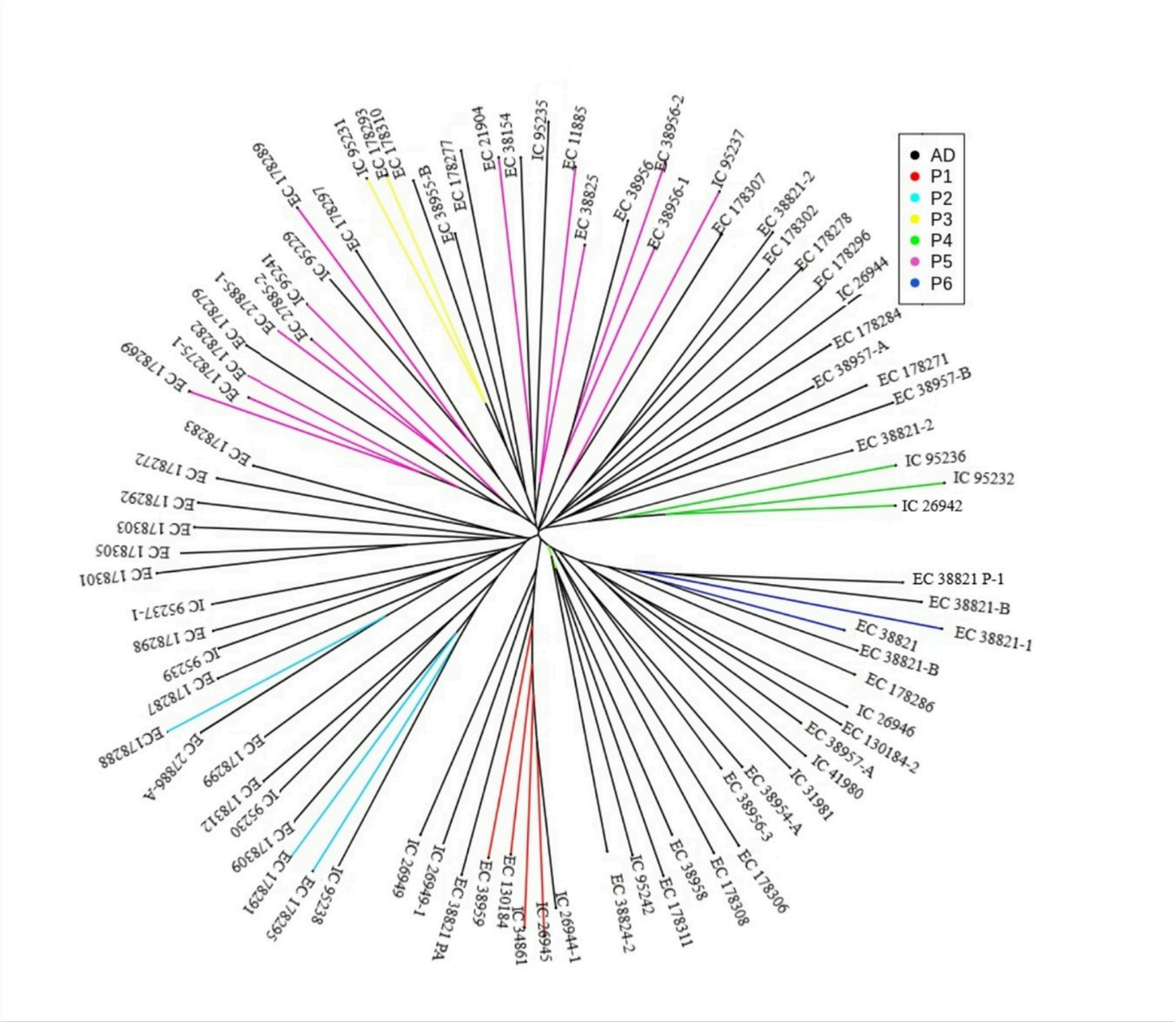

Figure 4

Phylogenetic relationships among 84 *Psophocarpus tetragonolobus* accessions.An unrooted neighbor-joining (NJ) tree was constructed using a pairwise P-distance matrix derived from 12,978 genome-wide SNPs filtered through the UNEAK pipeline. The matrix was generated in PowerMarker v3.25 and the tree were built using PHYLIP v3.697, then visualized using FigTree v1.4.3. The color-coded branches correspond to the six genetic subpopulations (K = 6) previously inferred by STRUCTURE analysis. Accessions with high subpopulation membership coefficients are grouped into distinct clades, while admixed individuals appear along intermediate branches. This topology reflects consistent genetic

differentiation and supports the STRUCTURE and PCA results, providing insight into the population stratification and evolutionary relatedness within the global *P. tetragonolobus* germplasm panel.

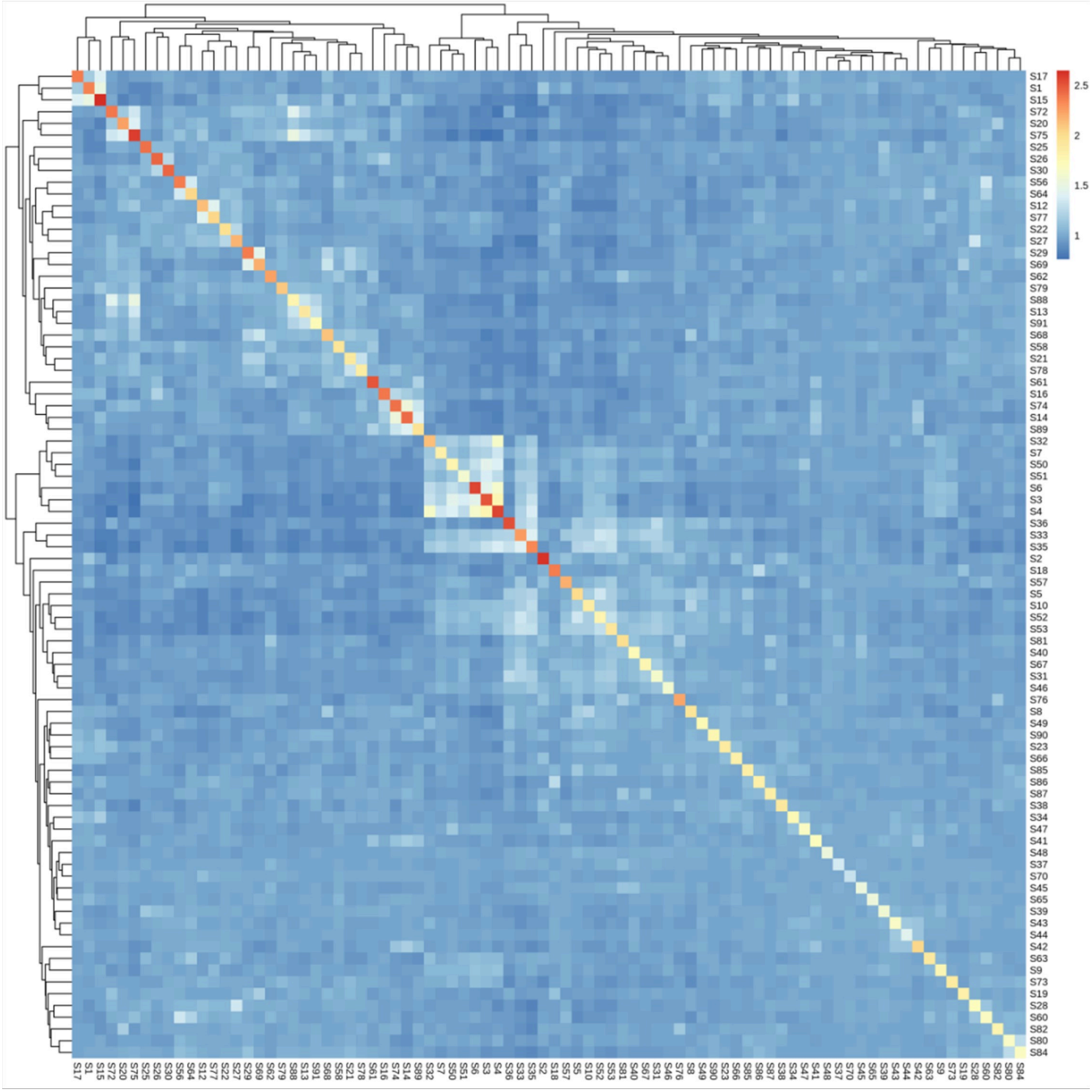

Figure 5

Kinship matrix of 84 *P. tetragonolobus* accessions. A heatmap showing pairwise relatedness among genotypes, generated using TASSEL5 from SNP data. Darker colors represent higher genetic similarity.

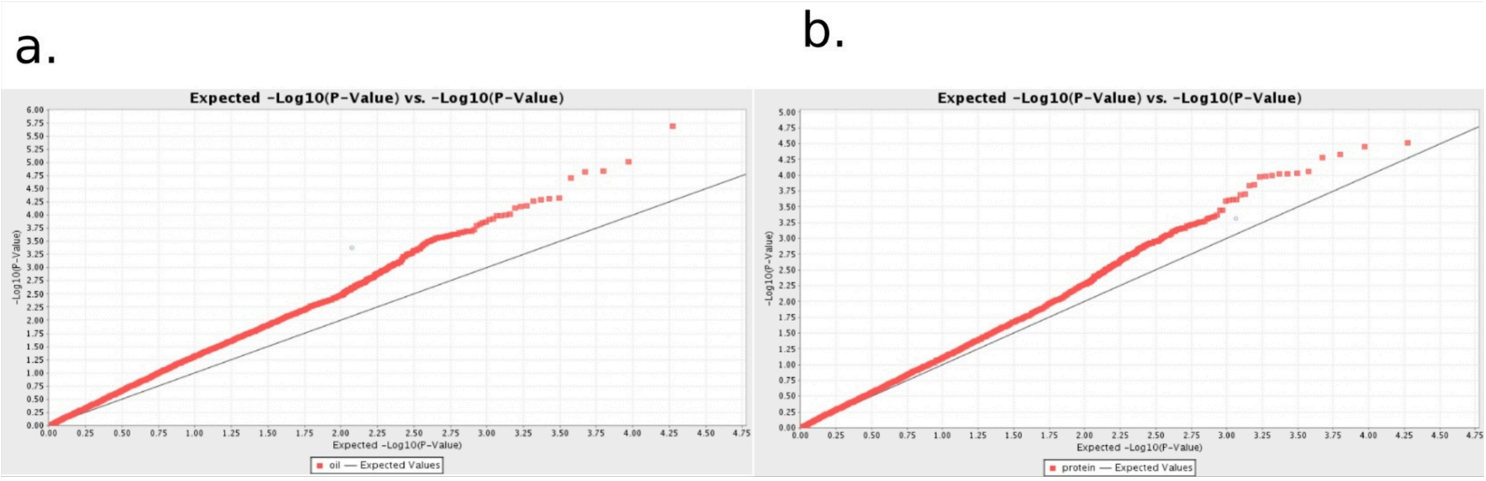

Figure 6

Quantile–quantile (Q–Q) plots of SNP-trait associations for oil and protein content in *Psophocarpus tetragonolobus*. (a) Q–Q plot of genome-wide SNPs associated with oil content. The observed $-\log_{10}$(p-values) are plotted against expected values under the null hypothesis of no association. Deviation of points from the diagonal line indicates potential SNPs significantly associated with oil trait variation. **(b)** Q–Q plot of genome-wide SNPs associated with protein content. The tail deviation reflects the presence of true marker-trait associations, while the rest of the distribution aligns well with the expected values, indicating good model control for population structure and relatedness.

Supplementary Files

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

- SupplementaryTable.docx