

# Genome-wide association and genomic prediction of seed nutritional traits and seed size in tepary bean (*Phaseolus acutifolius*) for low-input breeding

Sri Kiran Reddy

Texas A&M University

Benedict Analin

Texas A&M AgriLife Research

Vijay Joshi

[vijay.joshi@ag.tamu.edu](mailto:vijay.joshi@ag.tamu.edu)

Texas A&M University

---

## Research Article

**Keywords:** *Phaseolus acutifolius*, GWAS, genomic prediction, seed protein, essential amino acids, hundred-seed weight, seed width, low-input breeding

**Posted Date:** June 24th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10009866/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

**Version of Record:** A version of this preprint was published at Euphytica on October 14th, 2026. See the published version at <https://doi.org/10.1007/s10681-026-03821-7>.

# Abstract

Tepary bean (*Phaseolus acutifolius* A. Gray) is a heat- and drought-adapted legume with potential for low-input production but breeding resources for seed nutritional quality remain limited. A diversity panel of 206 accessions and four commercial checks was evaluated under certified organic field conditions in Uvalde, Texas, in 2024 for hundred-seed weight, seed width, total soluble seed protein, and nineteen free amino acids, and was genotyped with 49,384 high-quality SNPs generated by genotyping-by-sequencing. Genome-wide association analyses identified significant loci for protein percentage, seed size traits, and fifteen amino acid traits, with limited overlap among nutritional traits, indicating a largely modular genetic architecture. Essential amino acids of breeding interest, especially threonine, methionine, and lysine, showed distinct associations, while candidate genes near lead SNPs included a WNK kinase and 2-oxoglutarate/Fe(II)-dependent oxygenase for protein percentage, a vacuolar H<sup>+</sup>-ATPase-related region for hundred-seed weight, and an Aux/IAA-related region for seed width. Genomic prediction showed high predictive ability for seed size traits ( $r \approx 0.90\text{--}0.96$ ) and moderate predictive ability for protein and amino acid traits ( $r \approx 0.15\text{--}0.45$ ), supporting complementary use of marker-assisted and genomic selection.

These results provide the first genome-scale breeding benchmarks for seed nutritional traits in tepary bean under a low-input production environment. Because the experiment was conducted in a single environment and season, the identified loci should be treated as environment-specific hypotheses requiring validation across years and locations before routine deployment in breeding.

## Introduction

Tepary bean (*Phaseolus acutifolius* A. Gray) is native to the Sonoran Desert and has been cultivated for centuries by Indigenous communities of the southwestern United States and northern Mexico. Its domestication history reflects genomic adaptations to extreme heat and drought, as supported by a chromosome-scale reference genome (Moghaddam et al. 2021). Within the *Phaseolus* genus, tepary bean is closely related to common bean (*Phaseolus vulgaris*), enabling comparative genomics, interspecific gene transfer, and pre-breeding strategies for crop improvement (Porch et al. 2017). These attributes position tepary bean as a climate-smart, nitrogen-fixing legume with strong potential as a low-input food and ingredient crop in arid and semi-arid regions (López-Ibarra et al. 2021).

Tepary bean seeds are relatively rich in protein and dietary fiber and contain essential amino acids, suggesting a favorable nutritional profile compared with common bean (Porch et al. 2017; López-Ibarra et al. 2021). However, small seed size has limited broader adoption, as it influences water absorption, processing characteristics, and consumer acceptance (Porch et al. 2017; Giordani et al. 2022). Consequently, improving both seed size and protein quality represents a key objective for increasing the utility and marketability of tepary beans.

In *Phaseolus* and other legumes, association genetics provides a valuable framework for dissecting these complex traits. In common bean, image-based GWAS and integrative GWAS/RNA-seq studies have

identified polygenic control of seed shape and size (Giordani et al. 2022; Jurado et al. 2024). For nutritional traits, GWAS have mapped mineral content in common bean (Gunjača et al. 2021), while soybean studies have advanced understanding of amino acid regulation, revealing modular trait correlations and key regulatory loci (Qin et al. 2019). Parallel efforts in other pulses, including pea, chickpea, lentil, and cowpea, have identified loci influencing seed protein composition and nutritional quality (Warsame et al. 2024; Roorkiwal et al. 2022; Sari et al. 2024; Johnson et al. 2024; Chen et al. 2023).

The availability of a high-quality tepary bean reference genome reveals strong collinearity with *P. vulgaris* and signatures of domestication under heat stress, providing a foundation for genome-informed improvement. Research on tepary food ingredients supports the development of protein-rich flours but also highlights limitations in specific amino acids, particularly sulfur-containing amino acids, motivating efforts to optimize protein quantity and quality (López-Ibarra et al. 2021). Recent association studies in tepary have primarily relied on whole-genome resequencing and have focused on yield-related traits across environments, leaving the genetic basis of seed protein percentage, free amino acid composition, and their relationship with seed size relatively underexplored (Moghaddam et al. 2021; Ravelombola et al. 2024).

Genotyping-by-sequencing-based GWAS offers a cost-effective and scalable approach for SNP discovery and trait mapping in underutilized legume species. Multi-locus models, such as BLINK and FarmCPU, improve the detection of traits influenced by many genes, including those with complex architectures, such as seed composition traits. When anchored to the tepary reference genome, this approach enables hypothesis generation linking genomic regions to candidate genes involved in protein concentration, amino acid profiles, and seed development. These insights can inform both direct improvement of tepary bean and synteny-based translation to common bean.

In this study, a genome-wide association analysis of a diverse *Phaseolus acutifolius* panel was conducted to investigate total soluble seed protein content and free amino acid composition, with seed size traits used as complementary phenotypes. The objectives were to (i) quantify phenotypic variation in protein, amino acids, and seed size, (ii) identify genomic regions associated with nutritional and size traits, and (iii) evaluate the predictability of these traits using genomic selection models. Because seed nutritional and size traits in *Phaseolus* are sensitive to environmental conditions, the associations reported here reflect responses within a single certified-organic, low-input environment and therefore require multi-environment validation to assess locus stability and genotype × environment interactions.

## Materials and methods

### Plant material, growth conditions, and experimental design

A diversity panel of 206 *Phaseolus acutifolius* accessions obtained from the USDA National Plant Germplasm System, Western Regional Plant Introduction Station, was evaluated alongside four

commercial varieties used as replicated checks (CK1: Sacaton Brown; CK2: Sonoran White; CK3: Blue Speckled; CK4: Black), with seed sources listed in Supplementary Table S1. Field evaluations were conducted at the Texas A&M AgriLife Research and Extension Center in Uvalde, Texas, during a certified organic field season from July to October 2024. The trial adhered to standard organic practices, using drip irrigation as needed and OMRI-certified bioinsecticides for pest management. Accessions were sown on 25 July 2024 and harvested at physiological maturity between 26 and 30 September 2024.

The experiment was arranged as an augmented randomized complete block design with five blocks. In each block, all four checks were replicated to help control field variability and support the calculation of adjusted accession means for the unreplicated entries. Except for the checks, all phenotypic measurements for accessions represent a single experimental unit per accession, and repeated measurements were treated as technical subsampling or assay replication rather than as independent biological replicates.

## Phenotyping

### Seed size traits

At physiological maturity, pods were collected from each accession and commercial variety. For hundred-seed weight, three random 100-seed subsamples were taken from the harvested seed lot of each experimental unit, representing technical subsampling rather than independent biological replication, and seed weights were recorded using an ML54T/00 analytical balance with a precision of  $\pm 0.001$  g. Seed width was measured from scanned images of five fully dried seeds per accession and commercial variety, using ImageJ v1.54p and a scale bar for calibration. Seed width was retained as the image-derived seed-size trait used for downstream association analysis.

### Quantification of total soluble seed protein

Total soluble seed protein was estimated using a published method with minor modifications (Deans et al. 2018). Freeze-dried seeds were ground to a fine powder, and 15 mg of homogenized powder derived from the same harvested seed lot used for seed-size measurements was processed individually, suspended in 0.1 M NaOH, sonicated for 15 min, and centrifuged at  $14,000 \times g$  for 15 min. Protein concentration was determined using the Pierce Bradford Plus Protein Assay Reagent according to the manufacturer's instructions, with bovine serum albumin as an internal standard and absorbance read at 595 nm using a Varioskan LUX multimode microplate reader. Measurements were performed in triplicate and expressed as  $\text{mg g}^{-1}$  seed tissue and as percentage protein.

### Free amino acid extraction and analysis



Free amino acids were profiled following Joshi et al. (2019) with instrument-specific settings. Approximately 15 mg of seed powder from the pooled harvested seed material of the experimental unit was derivatized using the AccQ.Tag 3X Ultra-Fluor kit and quantified on a Waters ACQUITY H-Class UPLC coupled to an Xevo TQ mass spectrometer. Multiple reaction monitoring transitions, collision energies, and cone voltages were optimized in IntelliStart; data acquisition used MassLynx, and quantification used TargetLynx with external calibration.

## DNA extraction, GBS library construction, and SNP calling

Leaf DNA was extracted from 3- to 4-week-old seedlings using the DNeasy Plant Mini Kit according to the manufacturer's instructions, and DNA concentration was quantified with a DS-11 spectrophotometer. Library preparation and genotyping-by-sequencing were performed at the University of Minnesota Genomics Center using *ApeKI*. Reads were aligned to the *Phaseolus acutifolius* v1.0 reference genome, and SNPs were filtered before imputation to retain markers with minor allele frequency of at least 0.05, site missingness of 20% or less, and heterozygosity of 10% or less. After LD-kNNi imputation in TASSEL, 49,384 high-quality SNPs were retained for downstream analyses.

SNP density across the genome was determined by counting SNPs in non-overlapping 1 Mb segments along each chromosome, and unanchored scaffolds were combined into a single category (ChrUn) to simplify visualization of genome-wide SNP distribution. LD decay was analyzed using SNP data from the HapMap file, with pairwise LD ( $r^2$ ) calculated for SNPs within a 50 kb window on the same chromosome. For visualization,  $r^2$  values were binned by distance, and a smoothed decay curve was generated by cubic interpolation to identify the distance at which LD decayed to  $r^2 = 0.2$ .

Population structure was characterized by using both STRUCTURE and principal component analysis. After LD pruning in PLINK v1.9, 2,869 non-redundant SNPs were used for STRUCTURE 2.3.4 runs from  $K = 1$  to 10, and the optimal number of groups was identified using Structure Harvester and the Evanno method (Pritchard et al. 2000; Evanno et al. 2005). The first five principal components were included as covariates in GWAS models.

## Statistical analysis, GWAS, and genomic prediction

Data processing and statistics were performed using JMP and R v4.3.3, and figures were generated in R. GBS processing was performed by the University of Minnesota Genomics Center, and imputation used LD-kNNi in TASSEL version 5.2. Because accessions in the augmented randomized complete block design were unreplicated, BLUPs were obtained by borrowing strength from replicated checks and block structure, and adjusted entry values are reported as entry effects relative to the fitted model intercept; thus negative adjusted values reflect centering and do not imply negative biological trait values.

Stage-1 models produced adjusted entry values (BLUPs/BLUEs) and associated uncertainty measures, and the unweighted adjusted entry values were used as phenotypes in downstream GWAS and genomic

prediction analyses. Adjusted entry values were obtained using the augmentedRCBD package in R, following stage-wise analysis concepts described for genomic studies and unbalanced breeding data (Damesa et al. 2019; Endelman 2023; Fernández-González and Sánchez 2025; Schulz-Streeck et al. 2013). For each trait, the model form was:

$$y = \mu + \text{Block} + \text{Entry} + \epsilon$$

where  $y$  is the plot-level observation,  $\mu$  is the intercept, Block is the effect of the five blocks, Entry is the genotype or check effect, and  $\epsilon$  is the residual term. BLUPs were used for protein percentage, hundred-seed weight, and all amino acid traits, whereas BLUEs were used for seed width to report adjusted means on the observed scale.

GWAS was performed in GAPIT v3 using GLM, MLM, BLINK, and FarmCPU models, with population structure controlled by five principal components and kinship where applicable. A Bonferroni correction was used to declare genome-wide significance, and for the present panel of 206 accessions and 49,384 SNPs the threshold corresponded to approximately  $-\log_{10}(p) = 6$  (López-Hernández and Cortés 2019). Candidate genes were identified using a conservative  $\pm 10$  kb window around each significant SNP based on Phytozome annotations.

BLUP-based pseudo-heritability was estimated as the proportion of variance explained by genetic differences after adjustment for block effects and replicated checks. Genomic prediction was conducted with seven learners, BA, BB, BL, BRR, RF, rrBLUP, and SVM, using five-fold cross-validation and predictive ability defined as the Pearson correlation between adjusted phenotypes and genomic estimated breeding values. The genomic prediction framework followed widely used GBS and breeding-genomics approaches (Elshire et al. 2011; Federer and Crossa 2012; Möhring et al. 2014; Piepho et al. 2012; Smith et al. 2001; Williams et al. 2011).

## Results

### Phenotypic variation and nutritional diversity

Seed size and composition varied substantially across the tepary bean diversity panel. Hundred-seed weight ranged from 2 to 16 g per 100 seeds across the 206 accessions and four checks (Fig. 1A), and several accessions, such as PI 653254, PI 502217, PI 666351, and PI 535227, showed high seed weight values ranging from 14 to 16 g per 100 seeds. Checks showed an average of 10.5 g per 100 seeds.

Seed width ranged from 0.261 to 0.711 cm, with the checks clustering more narrowly around 0.536 to 0.640 cm (Fig. 1B). PI 502217, W6 38693, and PI 440805 showed the largest seed widths, ranging from 0.65 to 0.71 cm. Total soluble seed protein ranged from 10 to 38%, with an overall significant genotype effect detected by one-way ANOVA ( $p < 0.0001$ ) (Fig. 2). Checks averaged around 23% protein, and several accessions, including PI 201268, W6 38850, PI 640957, PI 440802, PI 440806, PI 200902, PI 321638, and PI 319439, showed comparatively high protein values.

Across all accessions, nineteen free amino acids were detected. Arginine (~ 28%), asparagine (~ 20%), aspartic acid (~ 18.5%), and glutamic acid (~ 12.4%) dominated the free amino acid pool (Fig. 3A). Principal component analysis revealed clear multivariate structure among amino acid traits, with PC1 (24.3%) driven primarily by essential and branched-chain amino acids, including methionine, histidine, leucine, isoleucine, and valine, and PC2 (13.0%) reflecting variation in nitrogen-rich amino acids and allantoin (Fig. 3B). Frequency distributions for all amino acids are provided in Supplementary Figure S3.

## Trait pseudo-heritability, SNP resources, and population structure

BLUP-based pseudo-heritability varied strongly across traits (Supplementary Table S4), reflecting considerable differences in genetic signal strength. Seed size traits and total soluble protein exhibited high pseudo-heritability, approximately 0.80 to 1.0, indicating strong genetic control under the single tested environment. Several amino acids, including arginine, asparagine, aspartic acid, glutamic acid, proline, and valine, also showed high values, whereas methionine, isoleucine, leucine, tyrosine, and glycine showed low pseudo-heritability ( $< 0.10$ ), indicating weaker genetic signal and greater environmental or residual variance.

GBS produced approximately 830 million raw reads, with a mean of approximately 3.95 million reads per sample. After alignment, quality filtering, and LD-kNNi imputation, 49,384 high-quality SNPs distributed across the eleven tepary chromosomes were retained for downstream analyses (Supplementary Figure S1). Population structure analysis identified two major subpopulations ( $K = 2$ ), consistent between STRUCTURE and PCA, with PC1 and PC2 explaining 23.2% and 7.3% of the variation, respectively (Supplementary Figure S4). Genome-wide LD decayed rapidly to  $r^2 = 0.2$  at approximately 14 kb (Supplementary Figure S2), supporting the use of a narrow  $\pm 10$  kb window for candidate-gene discovery.

## GWAS for total soluble protein and amino acid traits

Association analyses were performed on stage-1-adjusted phenotypes, with population structure controlled by five principal components and kinship, where appropriate. For total soluble protein content, four significant or suggestive associations were detected on chromosomes 1 and 8 (Fig. 4A). Three loci exceeded the Bonferroni threshold, S08\_8796692 ( $p = 1.38 \times 10^{-10}$ ), S01\_17593283 ( $p = 2.59 \times 10^{-8}$ ), and S08\_8784174 ( $p = 6.95 \times 10^{-7}$ ), whereas S08\_8761065 ( $p = 2.70 \times 10^{-6}$ ) was classified as suggestive. Candidate genes within  $\pm 10$  kb included *Phacu.CVR.008G093800* encoding a 2-oxoglutarate/Fe(II)-dependent oxygenase, *Phacu.CVR.001G119900* encoding a Ran-binding zinc finger, and *Phacu.CVR.008G093700* encoding a WNK kinase (Table 1; Supplementary Table S5).

Across the nineteen amino acid traits, fifteen showed at least one genome-wide significant association (Fig. 5; Supplementary Figures S5-S7). Representative examples included glutamine on chromosome 9 (S09\_37434888,  $p = 1.39 \times 10^{-19}$ ) and threonine on chromosome 8 (S08\_49548543,  $p = 6.95 \times 10^{-10}$ ). At the stringent Bonferroni threshold, overlap among amino acid associations was limited, supporting trait-

specific regulation rather than widespread pleiotropy, although shared signals may be underpowered and potentially more detectable under replicated multi-environment phenotyping, multi-trait models, or relaxed suggestive thresholds.

Candidate-gene analysis across all significant amino acid GWAS identified 192 unique genes involved in seed metabolism, nutrient transport, and developmental regulation (Table 1; Supplementary Table S5). Functional annotations showed enrichment for gene families involved in primary carbon and nitrogen metabolism and in membrane-associated and transport-related protein families, consistent with roles in amino acid transport and storage during seed development. Threonine revealed five significant SNPs and a plausible candidate region on chromosome 8 containing a Ser/Thr protein phosphatase family gene, while lysine displayed three significant associations, S07\_22884181, S11\_10209988, and S04\_6376564, near candidate genes including *Phacu.CVR.011G105200* encoding an S-adenosyl-L-methionine-dependent methyltransferase and *Phacu.CVR.004G052100* encoding WOX11.

## GWAS for hundred-seed weight and seed width

For hundred-seed weight, two genome-wide significant associations were detected, S02\_41997902 ( $p = 7.24 \times 10^{-8}$ ) and S01\_49536594 ( $p = 7.57 \times 10^{-7}$ ) (Fig. 4B; Table 1). Candidate genes included *Phacu.CVR.002G331300*, which encodes a vacuolar  $H^+$ -ATPase subunit involved in pH-dependent transport and endomembrane processes, and *Phacu.CVR.001G230400*, which encodes a CTR1-like protein kinase. For seed width, five significant associations were identified on chromosomes 1, 3, 5, 7, and 8 (Fig. 4C). Among the candidates, *Phacu.CVR.007G121900* encodes an auxin-responsive Aux/IAA family protein, consistent with known roles of auxin signaling in organ growth and seed development.

## Genomic prediction accuracy and breeding utility

Genomic prediction was used to quantify the predictability of seed nutritional and seed-size traits using the GBS marker set. Seven genomic prediction models were evaluated by five-fold cross-validation, with approximately 80% of accessions used for training and 20% held out for testing in each fold. Absolute predictive ability across traits is summarized in Fig. 6, while relative model performance benchmarked against rrBLUP is summarized in Supplementary Figure S8.

Seed-size traits showed consistently high predictability across models, with hundred-seed weight and seed width achieving predictive abilities of  $r \approx 0.90$ – $0.96$  for most learners. These results are consistent with previous reports indicating that morphological seed traits in pulses exhibit strong, polygenic genetic signals (Giordani et al. 2022; Warsame et al. 2024). In contrast, total soluble protein and free amino acid traits showed moderate predictive abilities ( $r \approx 0.15$ – $0.45$ ), mirroring the modular and mostly small-effect architectures revealed by GWAS and paralleling reports in soybean, chickpea, pea, and lentil (Qin et al. 2019; Roorkiwal et al. 2022; Sari et al. 2024; Johnson et al. 2024).

Across traits, rrBLUP and BRR performed comparably to Bayesian sparsity-based learners and nonlinear models, indicating that nutritional composition traits in tepary bean are governed largely by many small-effect loci rather than sparse large-effect signals. When averaged across traits, BRR showed a small

positive advantage relative to rrBLUP, whereas BA, BB, BL, RF, and SVM showed near-zero or slightly negative relative values. From a breeding perspective, these results support genomic selection as a practical strategy for seed-size improvement, while protein and amino acid improvement will likely require a combination of marker-assisted selection at prioritized loci and genomic selection to capture background polygenic variance.

## Discussion

# Seed protein and amino acids in tepary beans are controlled by modular, pathway-aware genetic loci

Prior studies on *Phaseolus* indicate significant genotype × environment interactions for seed traits, and it is expected that effect sizes and signal stability will differ across production systems. Since phenotyping was conducted in a single certified-organic environment over a single growing season, the GWAS associations reported here should be interpreted as environment-dependent hypotheses rather than universally stable loci. Within this context, the results reveal that total protein percentage and 15 of 19 free amino acids map to distinct genomic regions, supporting a modular genetic architecture rather than control by a single pleiotropic locus. These results align with legume GWAS in soybean, lentil, and pea, where protein quality traits and amino acid composition are dispersed across different genomic regions (Qin et al. 2019; Warsame et al. 2024; Johnson et al. 2024).

The association for protein percentage on chromosome 8 coincided with a WNK kinase and a 2-oxoglutarate/Fe(II)-dependent oxygenase, indicating a connection between regulatory signaling and the 2-oxoglutarate node that links carbon and nitrogen metabolism during seed filling, consistent with the central roles of GS/GOGAT, glutamate, and transamination in amino acid synthesis (Forde and Lea 2007; Saddhe et al. 2021). Collectively, these findings support a transport-and-sink model in which biosynthesis, long-distance transport, and sink import each contribute trait-specific variance that becomes detectable in different genomic windows during seed development (Tegeder and Rentsch 2010).

## Essential amino acid loci and nutritional improvement targets

Among essential amino acids, threonine and methionine are traditionally limiting in pulses and are key targets for improving dietary protein quality (Khazaei et al. 2019; Nosworthy et al. 2025). Threonine showed a significant genome-wide association on chromosome 8 within a phosphatase-rich region, providing a tepary-specific GWAS signal that motivates downstream validation in replicated multi-environment trials and/or segregating populations. This locus provides a prioritized candidate for marker development pending validation to increase threonine levels while monitoring related amino acids (Johnson et al. 2024; Forde and Lea 2007).

Methionine also showed genome-wide association, providing an important signal for this nutritionally critical amino acid in tepary bean. When considered jointly with the protein-percentage interval, this locus offers a potential strategy for improving protein quality without compromising total protein content, while genomic selection can capture residual polygenic variance distributed across the amino acid panel (Qin et al. 2019; Khazaei et al. 2019). In addition, lysine mapped to intervals including S11\_10209988 near a SAM-dependent methyltransferase (*Phacu.CVR.011G105200*) and S04\_6376564 near WOX11 (*Phacu.CVR.004G052100*), suggesting complementary regulatory mechanisms involving epigenetic control via the SAM cycle and developmental regulation via a WUSCHEL-related homeobox factor that integrates auxin/cytokinin cues (Lee et al. 2023; Meng et al. 2018; Zhao et al. 2009; Zhou et al. 2017).

Consistent with coordinated control in the aspartate-family pathway, feedback at key branch-point enzymes links Lys-Thr-Met pools, providing a mechanistic context for the observed lysine associations and their potential effects on related essential amino acids (Galili 1995; Azevedo et al. 1997; Goto et al. 2005; Jander and Joshi 2009). The chromosome 11 signal colocalizing with a SAM-dependent methyltransferase further connects lysine to the SAM cycle and epigenetic control during seed filling, because SAM supply via MAT enzymes is known to influence DNA and histone methylation and transcriptional programs in plants (Lee et al. 2023; Meng et al. 2018). Taken together, tepary bean's suitability for low-input, heat- and drought-adapted systems makes it a resilient legume protein source that can improve protein quality in resource-limited environments (Moghaddam et al. 2021; Barrera et al. 2024).

## Seed-size genetics and deployment under low inputs

Seed-size GWAS in tepary bean identified a small number of loci with modest effects, reflecting an overall polygenic architecture like that reported in image-based GWAS of common bean (Giordani et al. 2022). A chromosome 2 locus influencing hundred-seed weight was associated with a vacuolar H<sup>+</sup>-ATPase subunit and a 6-phosphogluconate dehydrogenase, while a chromosome 7 association for seed width localized near an Aux/IAA gene. The V-ATPase plays a key role in energizing tonoplast transport and maintaining vacuolar pH, which are vital for nutrient storage and cell expansion, thus linking endomembrane functions to organ growth (Krebs et al. 2010). Aux/IAA proteins are key repressors in the auxin regulatory pathway, and auxin influences cell division and expansion in developing seeds, placing the seed-width signal within a well-established growth control pathway (Leyser 2018; Wang and Estelle 2014).

These loci represent prioritized candidates for marker development and downstream validation rather than immediately deployable marker-assisted selection targets. Physical proximity does not establish causality, and the stability of marker-trait associations must be evaluated across environments and populations. Since tepary beans thrive in low-input environments and fix nitrogen biologically, even small increases in seed size could substantially boost marketability without requiring more fertilizer or irrigation (Moghaddam et al. 2021; Barrera et al. 2024). Compared with earlier tepary GWAS using whole-genome resequencing that focused on seed yield and hundred-seed weight across multiple

environments, the present GBS-based analysis in a single certified-organic environment identified a distinct chromosome 2 region enriched for growth-regulatory candidates, suggesting that tepary seed size is influenced by both environment-responsive loci and developmentally anchored loci (Giordani et al. 2022; Ravelombola et al. 2024).

## **Integrating genomic prediction with GWAS for seed composition and seed size**

Genomic prediction analyses complemented GWAS by quantifying the predictability of seed protein, amino acid, and seed-size traits across models of varying complexity. The high predictive ability observed for hundred-seed weight and seed width mirrors findings in common bean and pea, where seed morphological traits exhibit a strong polygenic signal that is readily captured by shrinkage-based models (Giordani et al. 2022; Warsame et al. 2024). These accuracies confirm that both major GWAS loci and background polygenic variance contribute consistently to seed-size traits.

In contrast, predictive abilities for total soluble protein and amino acid traits were moderate, mirroring the modular genetic architecture detected in GWAS. Similar patterns are reported in soybean and lentil, where amino acids and protein-quality traits show dispersed, low-effect loci and correspondingly modest prediction accuracies (Qin et al. 2019; Johnson et al. 2024). Across all traits, rrBLUP and BRR performed comparably to Bayesian and nonlinear learners, indicating that nutritional composition traits in tepary bean are governed largely by many small-effect loci rather than sparse large-effect signals. Improvements relative to rrBLUP were small and trait-consistent, with BRR showing slightly higher mean accuracy, a pattern consistent with reports across soybean, lentil, chickpea, and pea (Qin et al. 2019; Roorkiwal et al. 2022; Sari et al. 2024; Johnson et al. 2024).

Collectively, the GWAS-GP integration supports a dual breeding strategy: marker-assisted selection for the major protein-quality and essential-amino-acid loci and genomic selection to capture distributed polygenic variance influencing overall protein composition. Such MAS + GS pipelines are increasingly adopted in legume improvement and are well matched to tepary bean's adaptation to heat- and drought-stressed, low-input systems (Johnson et al. 2024; Moghaddam et al. 2021; López-Ibarra et al. 2021). Variation in BLUP-based pseudo-heritability across traits further contextualizes differences in GWAS power and prediction accuracy: traits with high pseudo-heritability showed stronger GWAS signals and higher predictive ability, whereas traits with low pseudo-heritability, such as methionine and some branched-chain amino acids, exhibited weaker signals, consistent with tighter metabolic regulation and greater environmental sensitivity.

## **Positioning tepary among pulses and delivering tools for introgression**

Within the expanding field of nutrition genetics across pulses, this study provides first-generation SNP markers for tepary bean targeting protein percentage, essential amino acids, and seed size (Qin et al. 2019; Warsame et al. 2024; Johnson et al. 2024; Chen et al. 2023). Coupled with a chromosome-scale

reference genome and intrinsic adaptation to heat and drought, these markers enable nutrition-focused breeding under low-input, nitrogen-fixing systems, an uncommon but increasingly valuable combination in climate-resilient agriculture (Moghaddam et al. 2021; López-Ibarra et al. 2021; Barrera et al. 2024). The identified loci facilitate marker-assisted selection for protein quality and seed size within tepary, while synteny can be leveraged to identify orthologous regions in common bean for comparative mapping and introgression. Genomic selection can further harness polygenic variation underlying amino acid composition, and integration with genome-environment pipelines developed for common bean and other legumes should aid deployment of these alleles in challenging environments (Qin et al. 2019; Chen et al. 2023; López-Hernández and Cortés 2019).

Comparative synteny studies among *Phaseolus* species should help identify orthologous variants in common bean and interspecific materials, and integrating selected SNPs or candidate loci identified here from tepary could accelerate the development of nutrition-focused ideotypes suited for arid conditions (Moghaddam et al. 2021; López-Ibarra et al. 2021; Giordani et al. 2022; López-Hernández and Cortés 2019; Forde and Lea 2007; Saddhe et al. 2021; Tegeder and Rentsch 2010; Barrera et al. 2024). Because all associations were identified in a single certified-organic environment, validation across multiple locations, seasons, and stress regimes is necessary to assess genotype × environment interactions and locus stability. The candidate genes reported here should therefore be viewed as systematic hypotheses rather than confirmed causal loci, and future work combining fine-mapping, multi-environment trials, transcriptomics, and metabolomics during seed filling will be needed to validate the roles of WNK signaling, 2-oxoglutarate oxygenases, amino acid transporters, and regulatory phosphatases.

## Conclusion

This study provides a genome-wide assessment of seed nutritional quality and seed size in tepary bean by integrating phenotyping, GWAS, and genomic prediction within a single low-input, certified-organic field environment. Using a diversity panel of 206 accessions, key genomic regions associated with total protein percentage, essential amino acids, and seed-size traits were identified. Trait-specific associations across fifteen amino acids support a modular genetic architecture rather than extensive pleiotropy, which is encouraging for marker stacking, index-based selection, and genomic selection. The threonine, methionine, and lysine signals represent especially relevant nutritional targets because they point to loci associated with essential amino acids that are often limiting in pulse-based diets. Likewise, hundred-seed weight and seed-width associations implicate biologically plausible growth-related candidates, including V-ATPase- and Aux/IAA-linked regions, that can be prioritized for follow-up validation. Together, these results provide a first set of marker-to-trait hypotheses, candidate genes, and genomic prediction benchmarks for improving both nutritional value and marketable seed size in tepary bean. More broadly, this work helps position tepary bean as a resilient legume protein source for hot, arid, and low-input production systems, while also creating a foundation for comparative mapping and eventual introgression-oriented work within *Phaseolus*. At the same time, the results should be interpreted cautiously because the study was conducted in a single environment and year, and because



limited seed quantity and prohibitive germination performance across parts of the panel currently constrain large-scale multi-season and multi-location validation trials. Accordingly, the reported loci should be regarded as breeding hypotheses and priority targets for future validation rather than as confirmed deployment-ready markers.

## Declarations

### Funding

This research was supported by the Sustainable Agriculture and Research Education (SARE) Graduate Student grant (GS24-299), Texas Department of Agriculture (TDA; GSC2023039, CDFA 10.170), and USDA-NIFA Hatch grant (project number 9647-1), awarded to Dr. Vijay Joshi.

### Author contributions

KR: Investigation, data curation, validation, visualization, formal analysis, software, writing-original draft, and editing; BA: formal analysis, software, writing-original draft, writing-review and editing; VJ: conceptualization, methodology, project administration, funding acquisition, resources, supervision, writing-review and editing.

### Conflict of interest

The authors declare no conflict of interest.

### Data availability

All sequencing data for tepary bean accessions have been submitted to the National Center for Biotechnology Information under the BioProject accession PRJNA1416895. Filtered GBS SNP data are available on FigShare at <https://doi.org/10.6084/m9.figshare.31158676>.

## References

1. Azevedo RA, Arruda P, Turner WL, Lea PJ (1997) The biosynthesis and metabolism of the aspartate derived amino acids in higher plants. *Phytochemistry* 46:395-419.
2. Barrera S, Berny Mier y Teran JC, Aparicio J, Diaz J, Leon R, Beebe S, et al. (2024) Identification of drought and heat tolerant tepary beans in a multi-environment trial study. *Crop Science* 64:3399-3416.
3. Chen Y, Xiong H, Ravelombola W, Bhattarai G, Barickman C, Alatawi I, et al. (2023) A genome-wide association study reveals region associated with seed protein content in cowpea. *Plants* 12:14.
4. Damesa T, Hartung J, Gowda M, Beyene Y, Das B, Semagn K, et al. (2019) Comparison of weighted and unweighted stage-wise analysis for genome-wide association studies and genomic selection. *Crop Science* 59.
5. Deans CA, Sword GA, Lenhart PA, Burkness E, Hutchison WD, Behmer ST (2018) Quantifying plant soluble protein and digestible carbohydrate content, using corn (*Zea mays*) as an exemplar. *Journal of Visualized Experiments* 138:58164.

6. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. (2011) A robust, simple genotyping-by-sequencing approach for high diversity species. *PLoS One* 6:e19379.
7. Endelman JB (2023) Fully efficient, two-stage analysis of multi-environment trials with directional dominance and multi-trait genomic selection. *Theoretical and Applied Genetics* 136:65.
8. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* 14:2611-2620.
9. Federer WT, Crossa J (2012) Screening experimental designs for quantitative trait loci, association mapping, genotype-by-environment interaction, and other investigations. *Frontiers in Physiology* 3:156.
10. Fernández-González J, Sánchez JI (2025) Optimizing fully-efficient two-stage models for genomic selection using open-source software. *Plant Methods* 21:9.
11. Forde BG, Lea PJ (2007) Glutamate in plants: metabolism, regulation, and signalling. *Journal of Experimental Botany* 58:2339-2358.
12. Galili G (1995) Regulation of lysine and threonine synthesis. *The Plant Cell* 7:899-906.
13. Giordani W, Gama HC, Chiorato AF, Garcia AAF, Vieira MLC (2022) Genome-wide association studies dissect the genetic architecture of seed shape and size in common bean. *G3: Genes, Genomes, Genetics* 12:jkac048.
14. Goto DB, Onouchi H, Naito S (2005) Dynamics of methionine biosynthesis. *Plant Biotechnology* 22:379-388.
15. Gunjača J, Carović-Stanko K, Lazarević B, Vidak M, Petek M, Liber Z, et al. (2021) Genome-wide association studies of mineral content in common bean. *Frontiers in Plant Science* 12.
16. Jander G, Joshi V (2009) Aspartate-derived amino acid biosynthesis in *Arabidopsis thaliana*. *The Arabidopsis Book* 7:e0121.
17. Johnson N, Boatwright JL, Bridges W, Thavarajah P, Kumar S, Thavarajah D (2024) Targeted improvement of plant-based protein: genome-wide association mapping of a lentil (*Lens culinaris* Medik.) diversity panel. *Plants, People, Planet* 6:640-655.
18. Joshi V, Joshi M, Silwal D, Noonan K, Rodriguez S, Penalosa A (2019) Systematized biosynthesis and catabolism regulate citrulline accumulation in watermelon. *Phytochemistry* 162:129-140.
19. Jurado M, García-Fernández C, Campa A, Ferreira JJ (2024) Identification of consistent QTL and candidate genes associated with seed traits in common bean by combining GWAS and RNA-Seq. *Theoretical and Applied Genetics* 137:143.
20. Khazaei H, Subedi M, Nickerson M, Martínez-Villaluenga C, Frias J, Vandenberg A (2019) Seed protein of lentils: current status, progress, and food applications. *Foods* 8:391.
21. Krebs M, Beyhl D, Görlich E, Al-Rasheid KAS, Marten I, Stierhof YD, et al. (2010) Arabidopsis V-ATPase activity at the tonoplast is required for efficient nutrient storage but not for sodium accumulation. *Proceedings of the National Academy of Sciences of the United States of America* 107:3251-3256.

22. Lee Y, Ren D, Jeon B, Liu H (2023) S-adenosylmethionine: more than just a methyl donor. *Natural Product Reports* 40:1521-1549.
23. Leyser O (2018) Auxin signaling. *Plant Physiology* 176:465-479.
24. López-Hernández F, Cortés AJ (2019) Last-generation genome-environment associations reveal the genetic basis of heat tolerance in common bean (*Phaseolus vulgaris* L.). *Frontiers in Genetics* 10:954.
25. López-Ibarra C, Ruiz-López FJ, Bautista-Villarreal M, Báez-González JG, Rodríguez Romero BA, González-Martínez BE, et al. (2021) Protein concentrates on tepary bean (*Phaseolus acutifolius* Gray) as a functional ingredient: in silico docking of tepary bean lectin to peroxisome proliferator-activated receptor gamma. *Frontiers in Nutrition* 8.
26. Meng J, Wang L, Wang J, Zhao X, Cheng J, Yu W, et al. (2018) METHIONINE ADENOSYLTRANSFERASE4 mediates DNA and histone methylation. *Plant Physiology* 177:652-670.
27. Moghaddam SM, Oladzad A, Koh C, Ramsay L, Hart JP, Mamidi S, et al. (2021) The tepary bean genome provides insight into evolution and domestication under heat stress. *Nature Communications* 12:2638.
28. Möhring J, Williams ER, Piepho HP (2014) Efficiency of augmented p-rep designs in multi-environment trials. *Theoretical and Applied Genetics* 127.
29. Nosworthy MG, Yu B, Zaharia LI, Medina G, Patterson N (2025) Pulse protein quality and derived bioactive peptides. *Frontiers in Plant Science* 16.
30. Piepho HP, Möhring J, Schulz-Streeck T, Ogutu JO (2012) A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal* 54:844-860.
31. Porch TG, Cichy K, Wang W, Brick M, Beaver JS, Santana-Morant D, et al. (2017) Nutritional composition and cooking characteristics of tepary bean (*Phaseolus acutifolius* Gray) in comparison with common bean (*Phaseolus vulgaris* L.). *Genetic Resources and Crop Evolution* 64:935-953.
32. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
33. Qin J, Shi A, Song Q, Li S, Wang F, Cao Y, et al. (2019) Genome wide association study and genomic selection of amino acid concentrations in soybean seeds. *Frontiers in Plant Science* 10:1445.
34. Ravelombola W, Manley A, Pham H, Brown M, Ruhl C, Ghosh P (2024) Genome-wide association study for seed yield of tepary bean using whole-genome resequencing. *International Journal of Molecular Sciences* 25:11302.
35. Roorkiwal M, Bhandari A, Barmukh R, Bajaj P, Valluri VK, Chitikineni A, et al. (2022) Genome-wide association mapping of nutritional traits for designing superior chickpea varieties. *Frontiers in Plant Science* 13.
36. Saddhe AA, Karle SB, Aftab T, Kumar K (2021) With no lysine kinases: the key regulatory networks and phytohormone cross talk in plant growth, development and stress response. *Plant Cell Reports* 40:2097-2109.

37. Sari H, Uhdre R, Wallace L, Coyne CJ, Bourland B, Zhang Z, et al. (2024) Genome-wide association study in chickpea (*Cicer arietinum* L.) for yield and nutritional components. *Euphytica* 220:84.
38. Schulz-Streeck T, Ogutu JO, Piepho HP (2013) Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and Applied Genetics* 126:69-82.
39. Smith A, Cullis B, Gilmour AR (2001) The analysis of crop variety evaluation data in Australia. *Australian & New Zealand Journal of Statistics* 43:129-145.
40. Tegeder M, Rentsch D (2010) Uptake and partitioning of amino acids and peptides. *Molecular Plant* 3:997-1011.
41. Wang R, Estelle M (2014) Diversity and specificity: auxin perception and signaling through the TIR1/AFB pathway. *Current Opinion in Plant Biology* 21:51-58.
42. Warsame AO, Balk J, Domoney C (2024) Identification of significant genome-wide associations and QTL underlying variation in seed protein composition in pea (*Pisum sativum* L.). *bioRxiv* 2024.07.04.602075.
43. Williams ER, Piepho HP, Whitaker D (2011) Augmented p-rep designs. *Biometrical Journal* 53:19-27.
44. Zhao Y, Hu Y, Dai M, Huang L, Zhou DX (2009) The WUSCHEL-related homeobox gene WOX11 is required to activate shoot-borne crown root development in rice. *The Plant Cell* 21:736-748.
45. Zhou S, Jiang W, Long F, Cheng S, Yang W, Zhao Y, et al. (2017) Rice homeodomain protein WOX11 recruits a histone acetyltransferase complex to establish programs of cell proliferation of crown root meristem. *The Plant Cell* 29:1088-1104.

## Figures

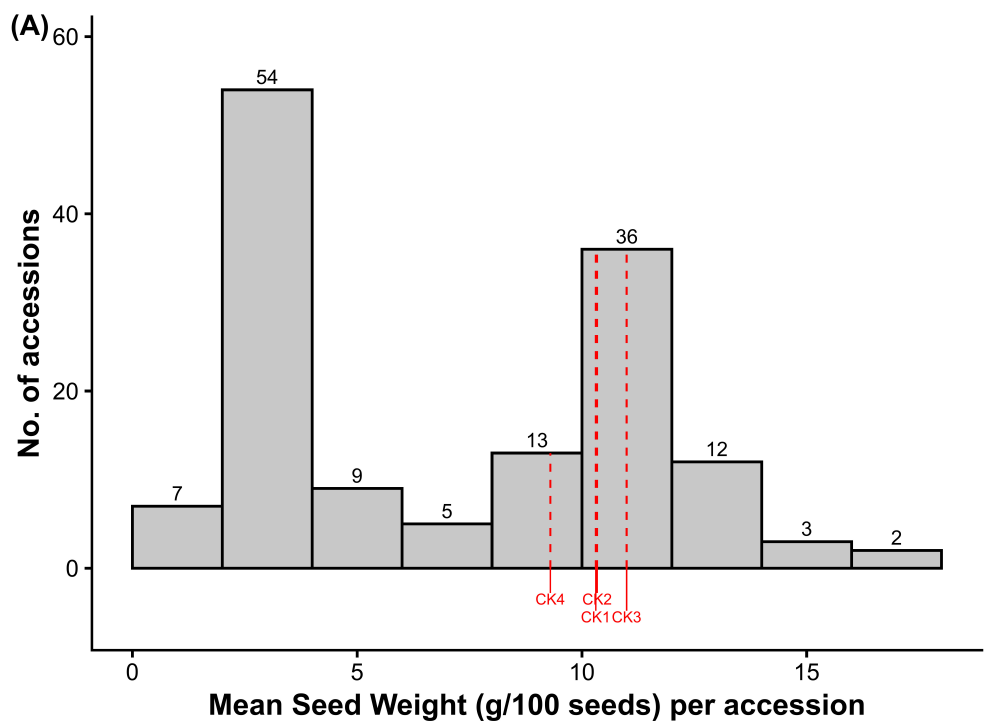
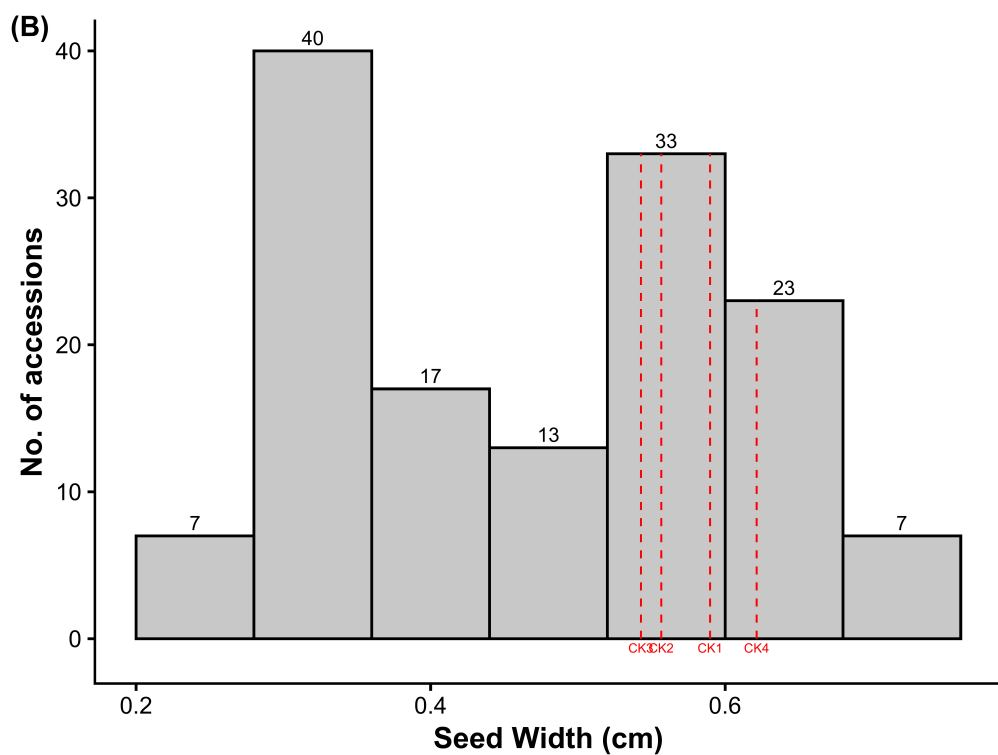


**Figure 1**

**Variation in seed size traits in the tepary bean diversity panel.** Frequency distributions of (A) hundred-seed weight and (B) seed width measured in 206 *Phaseolus acutifolius* accessions and four commercial checks grown under certified organic management at Uvalde, Texas.

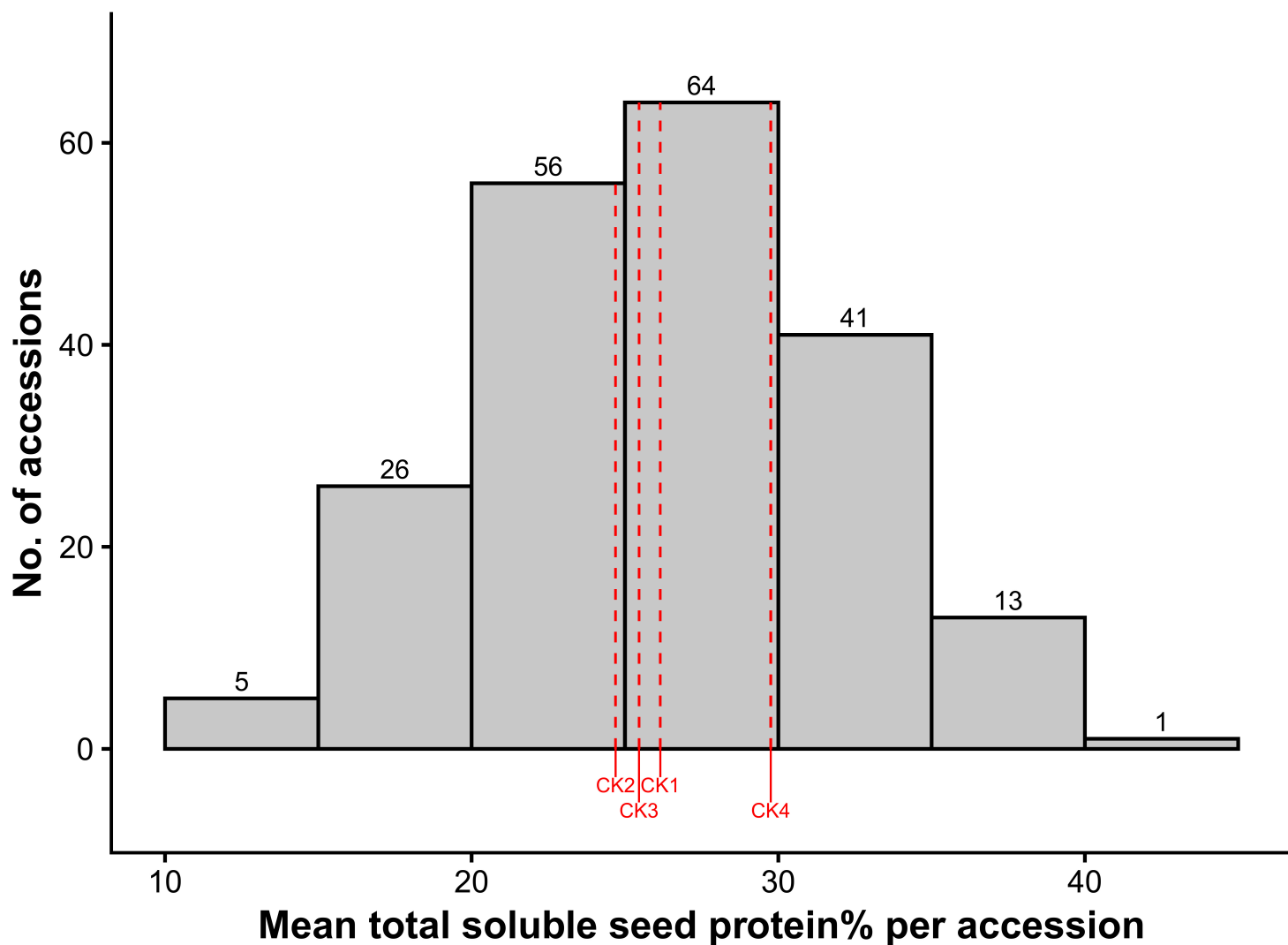


Figure 2

**Frequency distribution of total soluble seed protein in tepary bean.**Total soluble protein percentage estimated by Bradford assay for each accession and commercial check.

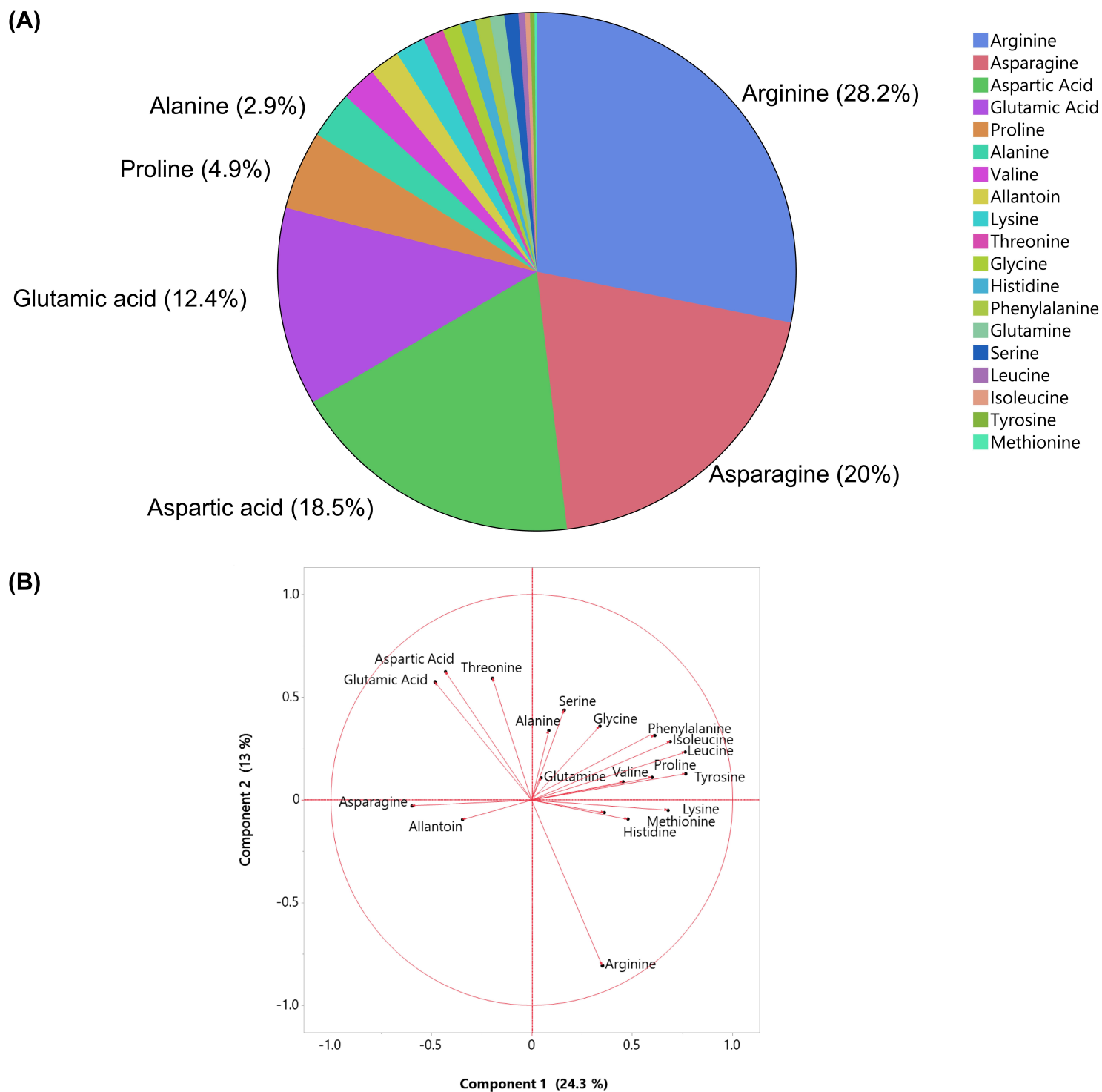


**Figure 3**

**Nutritional diversity of free amino acids in tepary bean seeds.** (A) Relative contribution of nineteen free amino acids to the total free amino acid pool and (B) principal component analysis of accession-level amino acid profiles.

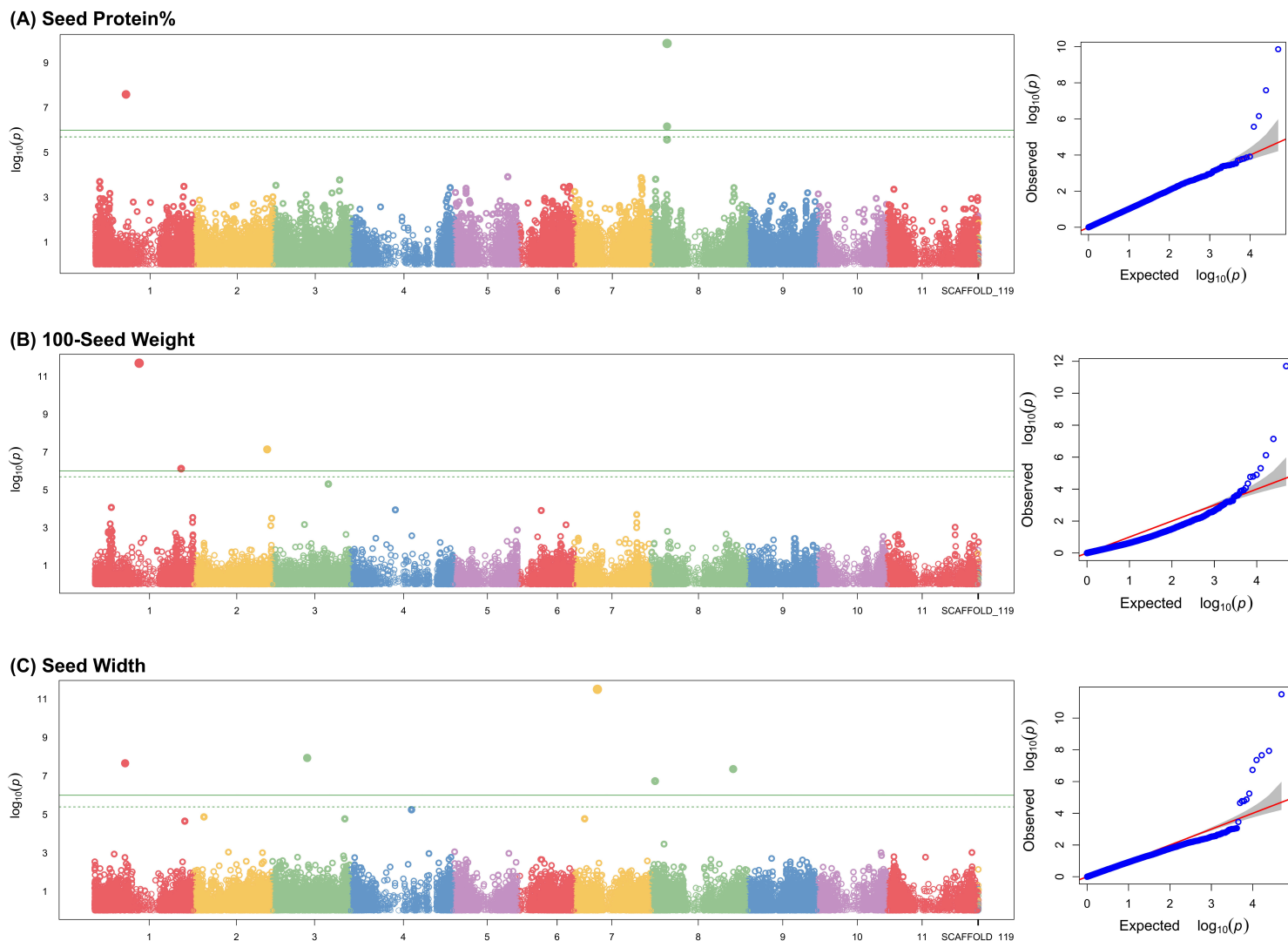


**Figure 4**

**Genome-wide association analysis of major breeding traits in tepary bean.** Manhattan and Q-Q plots for (A) total soluble protein, (B) hundred-seed weight, and (C) seed width. The dashed line indicates the Bonferroni threshold based on 49,384 SNPs.



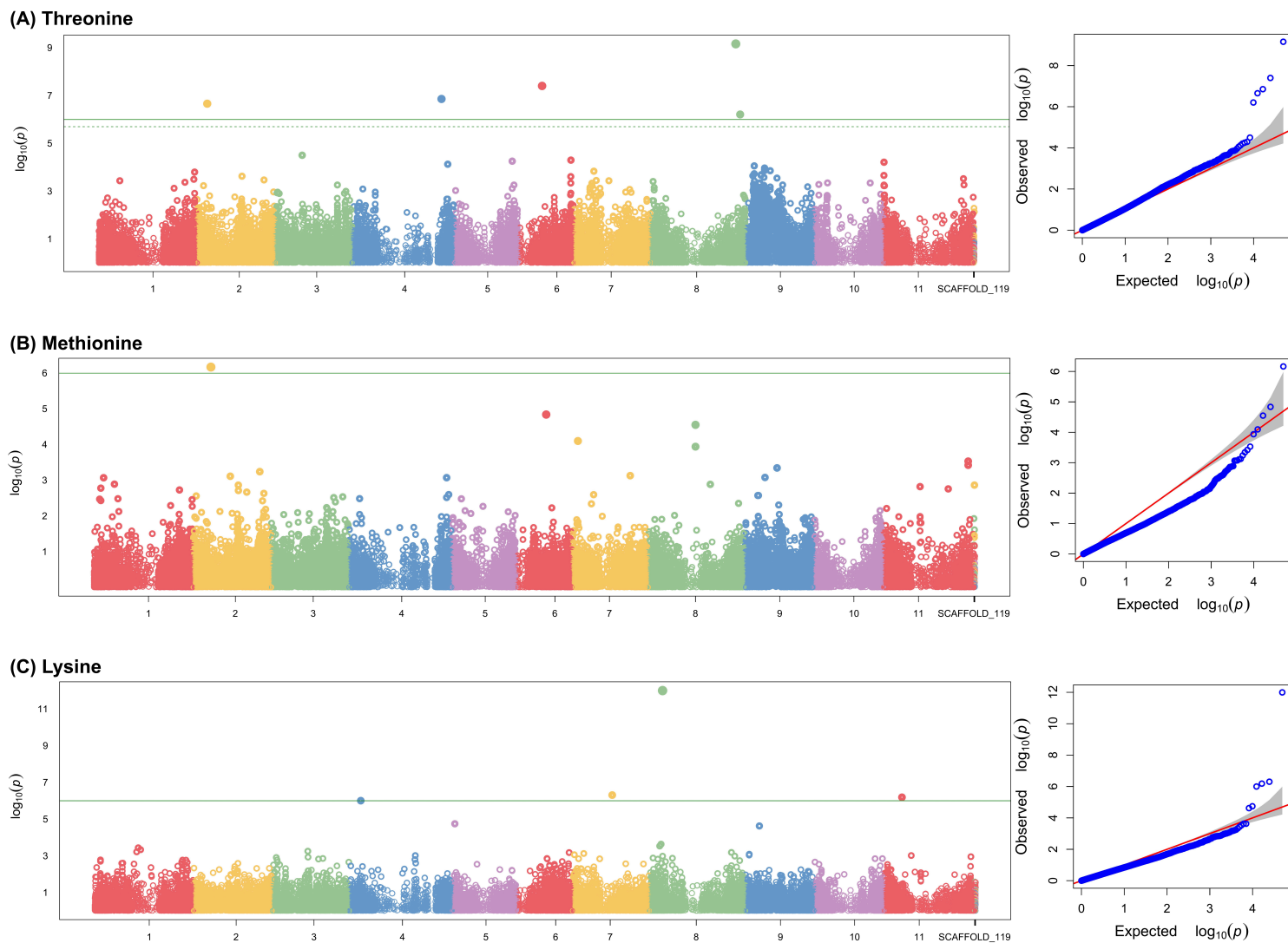


**Figure 5**

**Association signals for breeding-priority essential amino acids.** Manhattan and Q-Q plots for (A) threonine, (B) methionine, and (C) lysine. Lead SNPs and nearby candidate genes are summarized in Table 1 and Supplementary Table S5.

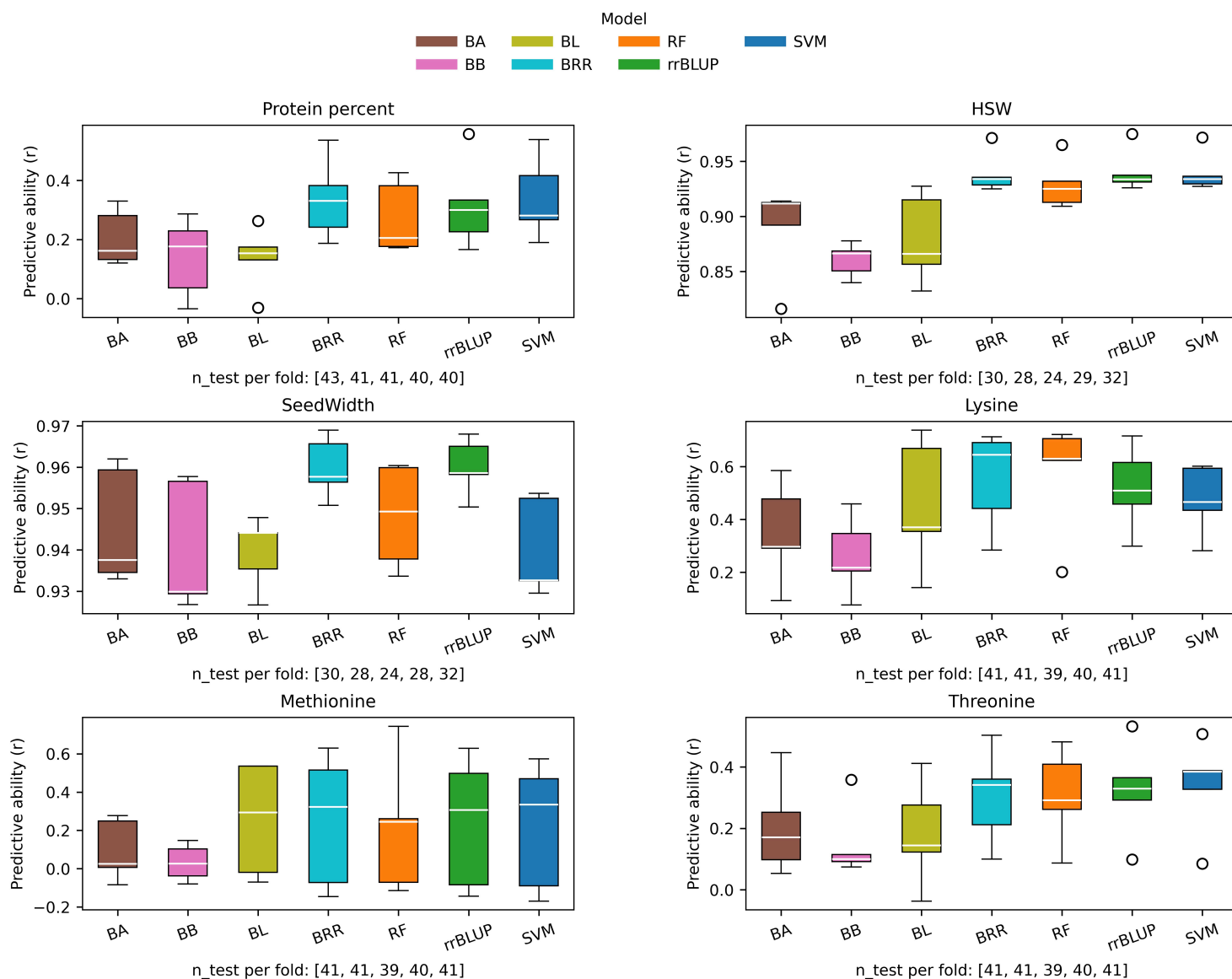


**Figure 6**

**Predictive ability of genomic prediction models for seed size and nutritional traits.** Predictive ability was estimated as the Pearson correlation between adjusted phenotypes and genomic estimated breeding values using five-fold cross-validation across seven genomic prediction models.

## Supplementary Files

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

- [AdditionalFile2.pdf](#)
- [AdditionalFile1.xlsx](#)
- [AdditionalFile7.xlsx](#)
- [AdditionalFile4.xlsx](#)

- [AdditionalFile5.pdf](#)
- [AdditionalFile6.pdf](#)
- [AdditionalFile3.xlsx](#)
- [AdditionalFile13.jpeg](#)
- [AdditionalFile9.jpg](#)
- [AdditionalFile10.jpg](#)
- [AdditionalFile11.jpg](#)
- [AdditionalFile12.jpg](#)
- [AdditionalFile8.jpg](#)
- [AdditionalFile11.jpg](#)
- [AdditionalFile13.jpeg](#)
- [Supplementarymaterial.docx](#)