Received: 29 April 2023 | Revised: 27 October 2023 | Accepted: 8 November 2023

DOI: 10.1002/ppp3.10470

Plants People Planet PPP Open Access

RESEARCH ARTICLE

Targeted improvement of plant-based protein: Genome-wide association mapping of a lentil (*Lens culinaris* Medik.) diversity panel

Nathan Johnson [1] | J. Lucas Boatwright [1,2] | William Bridges [3] | Pushparajah Thavarajah [1] | Shiv Kumar [4] | Dil Thavarajah [1]

[1]Plant and Environmental Sciences, Clemson University, Clemson, South Carolina, USA

[2]Advanced Plant Technology, Clemson University, Clemson, South Carolina, USA

[3]Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA

[4]International Centre for Agricultural Research in the Dry Areas (ICARDA), Rabat-Institute, Rabat, Morocco

Correspondence
Dil Thavarajah, Plant and Environmental Sciences, Clemson University, 113 Biosystems Research Complex, Clemson, SC, USA.
Email: dthavar@clemson.edu

Funding information
FoodShot Global; SC Department of Agriculture; Good Food Institute; United States Department of Agriculture; National Institute of Food and Agriculture, Grant/Award Numbers: 2018-51300-28431, 2021-51300-34805; Department of Agriculture

Societal Impact Statement

The world is increasingly looking to plant-based sources to meet its protein needs. Multiple factors are driving this progression, ranging from nutritional and ethical considerations to climate change and population growth. As a pulse crop, lentil is ideal to help meet this change in demand. However, plant-based proteins have limiting amino acids and lower protein digestibility compared to animal-based proteins. This research identifies genetic markers that can be used to accelerate breeding of protein quality traits in lentil to ultimately help meet the rising demand in high-quality plant-based protein and bolster global food and nutritional security.

Summary

- Lentil (*Lens culinaris* Medik.) contains ~25% high-quality protein in addition to high concentrations of prebiotic carbohydrates and micronutrients, such as folate, iron, zinc, and selenium. As animal-based protein's economic and environmental costs rise, plant-based protein sources, such as lentil, will become increasingly important to global food systems. Consequently, evaluating and targeting protein quality traits for genomic-assisted breeding is a valuable objective for lentil breeding programs.
- A diversity panel of 183 breeding lines was analyzed for protein quality traits, including amino acids and protein digestibility. Genotyping-by-sequencing (GBS) data were used to assess population structure and conduct genome-wide association studies (GWAS). Genes in local linkage disequilibrium (LD) with significant single nucleotide polymorphism (SNP) markers were identified and categorized by homology.
- Protein quality traits showed a wide range of variation. Repeatability estimates were low to moderate across traits. Twelve traits were strongly correlated with each other ($r > .7$). Admixture analysis identified six ancestral subpopulations, which also demonstrated clustering in principal component analysis. Ten different traits had significant SNP associations; two loci were shared across multiple traits. Twenty-seven candidate genes, including glutathione S-transferase, protease family, and gibberellin 2-beta-dioxygenase genes, were identified.

- This paper identifies SNP markers associated with lentil protein quality traits. Once validated, these SNPs could accelerate lentil protein quality breeding efforts. By targeting lentil's limiting amino acids (methionine and cysteine) and protein digestibility through marker-assisted selection, the nutritional value of lentil's protein content could be increased without the need to alter total protein content.

KEYWORDS
admixture, amino acids, breeding, genome-wide association study, lentil, protein, protein digestibility, pulses

1 | INTRODUCTION

Cereal crops have experienced dramatic increases in yields since the Green Revolution (Evans & Lawson, 2020). These increases have allowed growth in the global population, which is projected to further increase to 9.8 billion in 2050. Cereal crops have low concentrations of dietary protein and other micronutrients. Dietary protein rich in essential amino acids (AAs) must be supplied through other food sources, including animal- or plant-based proteins. Animal-based proteins provide the highest quality protein for human nutrition because of their higher protein digestibility (PDg) and essential AA concentrations. However, animal-based protein sources require higher production inputs (e.g., water, energy, and animal welfare costs) than plant-based protein sources. Globally, these requirements constrain access to animal protein because of increased production costs and environmental sustainability concerns. Legumes, such as lentils (*Lens culinaris* Medik.), can serve as adequate protein sources at a lower cost, with additional agricultural benefits such as nitrogen fixation when used as a rotation crop. The plant-based protein market has been steadily growing. Globally, the plant-based protein market will continue to expand and will reach $28 billion in 2022. North America is the largest region contributing to this increasing demand, representing 28–30% of the global protein market in 2022 (GFI, 2023). To meet the global demand for plant-based protein, ingredient suppliers have expanded the need for novel plant proteins from legumes. Pulse crops, including lentils, are the most in-demand ingredient for this industry because of their high protein content with low fat.

Protein quality, in addition to protein concentration, is an important consideration when evaluating plant-based protein sources. Lentil is a staple part of many traditional diets worldwide, features prominently in Mediterranean diets, and lowers the risk of all-cause, cardiovascular, and cancer mortality (Papandreou et al., 2019). A 100 g serving of lentil provides 25 g of protein or 50% of the recommended dietary allowance (U.S. Department of Agriculture, 2019; National Research Council Subcommittee on the Tenth Edition of the Recommended Dietary Allowances, 1989) and is a rich source of prebiotic carbohydrates, vitamins, and minerals (Johnson et al., 2020). As many people transition away from animal-based protein sources, lentil is an excellent alternative from nutritional, economical, and environmental perspectives. Lentil is less expensive to purchase than meat, reduces the impact of emissions on the environment, and lowers nitrogen fertilizer requirements by fixing atmospheric nitrogen in root nodules (Foyer et al., 2016; Semba et al., 2021). However, a challenge faced by lentil protein, along with other plant-based proteins, is that the protein is of lower quality than animal protein.

Proteins are macromolecules composed of AAs bound together by peptide bonds. A healthy diet requires a sufficient quantity of protein and sufficient quantities of the essential AAs phenylalanine (Phe), valine (Val), tryptophan (Trp), threonine (Thr), isoleucine (Ile), methionine (Met), histidine (His), leucine (Leu), and lysine (Lys). These are AAs that humans cannot synthesize and, therefore, must consume in their diets or suffer malnutrition. Lentil is a good source of aspartate (Asp), alanine (Ala), arginine (Arg), and glutamate (Glu); however, lentils have limited quantities of the sulfur-containing AAs (S-AAs), Met, and Cys (Salaria et al., 2022). The body can synthesize Cys from Met; consequently, they are combined for intake requirements. Cys is vital in protein folding because of its ability to form disulfide bonds; Met is significant for its role as the AA that begins translation as well for its derivatives, glutathione, and S-adenosyl methionine (SAM), which are important in oxidative protection and DNA methylation, respectively, and have been examined in such diverse pathologies as obesity and Parkinson's disease (Barbosa et al., 2021; Jalgaonkar et al., 2022). In addition to AA composition, protein quality also depends on PDg, which determines how well the protein can be catabolized during digestion and utilized by the body. Lentil PDg is ~84%, which is excellent relative to other crops (cf. oat 72%, wheat 77%, and soybean 78%); however, plant-based proteins tend to have a lower PDg than animal-based proteins (cf. meat/poultry/fish 94%, milk 95%, and egg 97%) (Gilbert et al., 2011).

To ensure nutritional food security, lentil biofortification aims to improve protein quality and other nutritional traits (Kumar et al., 2016). Significant genetic variation has been observed for protein traits in lentil, including protein content, storage protein structure and weight, and AA concentrations (Alghamdi et al., 2014; Ghumman et al., 2019; Hang et al., 2022). Accelerating the rate of genetic gain by reducing breeding cycle time is a primary objective in breeding programs, and genomic tools have allowed for significant gains through genetic markers and molecular breeding approaches (Cobb et al., 2019; Kumar et al., 2021). Genome-wide association studies (GWAS) have been extensively used in plant science to identify genetic markers and candidate genes associated with a range of traits (Tibbs Cortes et al., 2021). In lentil, GWAS has been used to identify

markers associated with days to flower, seeds per pod, 100 seed weight, salinity tolerance, *Aphanomyces* root rot resistance, iron (Fe) and zinc (Zn) concentrations, prebiotic carbohydrate concentrations, and AA concentrations quantified by near-infrared spectroscopy (Dissanayake et al., 2021; Hang, 2021; Hang et al., 2022; Johnson et al., 2021; Kumar et al., 2019; Ma et al., 2020; Rajendran et al., 2021). Protein-quality traits have also been explored via GWAS in cereal crops (Chen et al., 2016) and legumes such as chickpea (Karaca et al., 2019) and common bean (Katuuramu et al., 2018). Genomic markers can be used to accelerate breeding efforts. One prominent method is marker-assisted selection, which leverages genetic markers associated with a causal allele or gene to select for a trait, such as through introgression and backcrossing. This strategy has been successfully used to develop maize with high beta-carotene content (Muthusamy et al., 2014) and rice with increased flood resistance (Bailey-Serres et al., 2010).

Population structure and kinship between individuals can produce linkage disequilibrium (LD) between unlinked loci (Mangin et al., 2012). This is particularly concerning in GWAS, potentially increasing the rate of false positives, which associate a marker with a phenotype despite the marker having no true significant effect on the trait. This has been shown in multiple crops including the legume, common bean (Diniz et al., 2019), and is a concern in lentil. Several genome-wide association models have been developed to address this concern (Wang & Zhang, 2021), and many of which were evaluated for application in this study. Among these is the multiple locus mixed linear model (MLMM) (Segura et al., 2012). MLMM incorporates principal components as fixed effects to help address population structure and kinship estimates as random effects to help address relatedness between individuals. In addition, a forward–backward stepwise regression is used to include associated markers as cofactors. Consequently, the risk of false positives is reduced, and power is increased. Quantile–Quantile plots can help evaluate how well a model accounts for population structure and kinship and whether these are systematically contributing to false positives.

Improving lentil protein quality through biofortification is a desirable breeding objective; however, genetic markers and genes associated with the relevant traits are limited. Therefore, the objectives of this study were to (1) quantify protein quality traits (AAs and PDg) in a lentil association mapping population grown under greenhouse conditions, (2) evaluate ancestral subpopulation global distribution and subpopulation effects on protein quality traits, and (3) identify single nucleotide polymorphism (SNP) markers and candidate genes associated with these traits.

2 | MATERIALS AND METHODS

2.1 | Diversity panel composition

Two mapping populations were obtained from the International Center for Agricultural Research in the Dry Areas (ICARDA) (Table S1). The Heat Tolerance Mapping Population was designed to explore heat and drought stress tolerance in lentil. The Global Mapping Population is a diversity panel designed to incorporate global lentil diversity. These were grown with two replicate pots per accession in the Clemson greenhouse complex in 2018 (Johnson et al., 2021). Samples from the two populations were combined for analysis. These samples originated from 55 different countries. After accounting for population overlap and low yields from some accessions and replicates, 183 unique accessions with 1 to 4 replicates each were analyzed for protein quality traits.

2.2 | AA analysis

Reagents, solvents, and high-purity standards for AA analysis were purchased from Sigma–Aldrich Co. (St. Louis, MO), Thermo Fisher Scientific (Waltham, MA), and VWR International (Radnor, PA). Ultrapure water was used in all analyses (PURELAB flex 2 system, ELGA LabWater North America, Woodridge, IL). The AA analysis is reported elsewhere (Madurapperumage et al., 2022) as an adaptation from the literature (Gehrke et al., 1985; Manneberg et al., 1995). In brief, 40 mg per sample of lentil flour was used in the analysis. A pre-hydrolysis oxidation step using 5 mL of performic acid (16-h refrigeration) was included to convert Cys and Met to their more stable derivatives, methionine sulfone and cysteic acid. Samples were vacuum evaporated to dryness before 4.9 mL of 6 M hydrochloric acid (HCl) and 0.1 mL internal standard mix (norvaline and sarcosine) were added. A gravity convection oven at 110°C for 24 h was used to hydrolyze peptide bonds, followed by evaporation to dryness of a subsample and rehydration with HPLC mobile phase. AA analysis was performed via reverse phase high-performance chromatography (HPLC) on a 1100 series Agilent system (Agilent Technologies, Santa Clara, CA, USA) according to Madurapperumage et al. (2022), which is a method adapted from Agilent application notes (Agilent Application Note, 2010; Long, 2015). AAs were separated with an Agilent Poroshell HPH-C18 analytical column (3 × 100 mm) in series with the corresponding guard column (3 × 5 mm) and detected with a diode array detector at 338 and 262 nm wavelengths. A gradient method was employed with online sample derivatization with *o*-phthalaldehyde (OPA) and 9-fluorenylmethyl chloroformate (FMOC). A lab reference lentil sample was included in every digestion batch to monitor batch-to-batch variation, and an AA standard mix was run on the HPLC instrument before analyzing each batch of samples. Calibration standards (9–900 pmol/μL) with internal standards norvaline and sarcosine (500 pmol/μL) were run, and linear calibration models were generated based on peak areas for calculating sample AA concentrations, which were converted into percent of lentil flour. Total amino acid (TA) concentration was calculated by summing all AA concentrations for each sample. The percent of total AA concentration was calculated for each AA. Consequently, the 17 AA concentrations resulted in 35 AA traits—17 AA, 17 AA percent of total AA (AA:TA), and TA concentration (Table S2).

2.3 | In vitro protein digestibility analysis

PDg was measured using the Megazyme Protein Digestibility Amino Acid Score assay kit with a modified protocol for a 100 mg sample size (Megazyme, 2019). The protocol was followed precisely except all masses and volumes were divided by 5. Because of expected underestimation of some AAs caused by acid hydrolysis, reference AA values from the U.S. Department of Agriculture Food Data Central (U.S. Department of Agriculture, 2019) were used for the PDg calculations; these included Proline (Pro, 1.03%), Lys (1.72%), His (0.69%), and Arg (1.90%). The Megazyme Excel calculator was modified to change the approximate sample mass from 0.5 to 0.1 g. In addition to the controls included in the assay kit, a lab reference lentil sample was included in every batch to monitor batch-to-batch variation.

2.4 | Summary statistics and correlations

Protein quality trait means, standard deviations, and ranges were calculated for each accession in JMP 14.0.0 (Tables 1 and S3). Histograms of trait distributions were fit with density curves for the normal distribution using estimates of the mean and standard deviation (Figure S1). Percent recommended dietary allowance (RDA) estimates were calculated for the essential AAs His, Ile, Leu, Lys, Met, Phe, Thr, Trp, and Val as well as for total AA concentration. Estimates were for a 72-kg adult consuming 100 mg of lentil (15% moisture content) per day given the following dietary requirements: 8–12 mg/kg His, 10 mg/kg Ile, 14 mg/kg Leu, 12 mg/kg Lys, 13 mg/kg Met + Cys, 14 mg/kg Phe + Tyr, 10 mg/kg Val, and 0.8 g/kg protein (National Research Council Subcommittee on the Tenth Edition of the Recommended Dietary Allowances, 1989). To estimate repeatability, a model was developed with trait concentration as the response variable and genotype as a random effect. Repeatability is the proportion of phenotypic variance attributable to genetic variance and provides an upper bound to broad-sense heritability (H^2). Repeatability equals H^2 when all differences between genotypes are assumed to be genetic (Kruijer et al., 2014). Pearson's correlation coefficients (r) were calculated in JMP for protein quality traits using accession means across replicates (Table 2).

2.5 | Genome-wide association studies

A previously generated variant call format (VCF) file was used for genetic analyses (Johnson et al., 2021). In brief, the TASSEL-GBS pipeline (Glaubitz et al., 2014) was used to process raw genotyping-

TABLE 1 Mean concentration, concentration range, repeatability estimates, and %RDA for lentil protein quality traits.

Trait	Mean (%) ± SD	Range (%)	%RDA[a]	Repeatability %[b]
Ala	1.24 ± 0.11	0.77–1.55	N/A	20.0
Arg	2.72 ± 0.33	1.53–3.62	N/A	19.5
Asp	3.87 ± 0.48	2.29–6.04	N/A	22.2
Cys	0.22 ± 0.03	0.15–0.29	20	18.8
Glu	4.45 ± 0.40	2.74–5.63	N/A	19.7
Gly	1.25 ± 0.11	0.81–1.56	N/A	19.3
His	0.61 ± 0.10	0.18–1.00	60–90	14.2
H-Pro	3.16 ± 0.95	1.5–6.01	N/A	11.6
Ile	1.24 ± 0.11	0.77–1.52	100	21.5
Leu	2.25 ± 0.20	1.37–2.8	100	21.4
Lys	1.33 ± 0.23	0.56–1.89	100	19.5
Met	0.21 ± 0.02	0.11–0.27	19	26.4
Phe	1.29 ± 0.23	0.39–1.72	100	3.8
Pro	2.49 ± 0.49	1.44–3.96	N/A	18.1
Ser	1.40 ± 0.13	0.85–1.76	N/A	20.7
Thr	1.05 ± 0.10	0.6–1.3	100	22.4
Val	0.21 ± 0.02	0.14–0.26	25	18.2
TA	28.99 ± 2.90	18.27–36.05	43	23.5
PDg	88 ± 1	86–91	N/A	34.9

Abbreviations: Ala, alanine; Arg, arginine; Asp, aspartate; Cys, cystine; Glu, glutamate; Gly, glycine; His, histidine; H-Pro, hydroxyproline; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; N/A, not applicable; PDg, protein digestibility; Phe, phenylalanine; Pro, proline; SD, standard deviation; Ser, serine; TA, total amino acid; Thr, threonine; Val, valine.

[a]Percent recommended dietary allowance (%RDA) estimates were calculated for a 72-kg adult consuming a 100 g serving of lentil (15% moisture content) per day.

[b]Repeatability is the proportion of phenotypic variance attributable to genetic variance and provides an upper bound to broad-sense heritability (H^2) (Kruijer et al., 2014).

TABLE 2 Correlations between protein quality traits grouped by amino acid family of lentil diversity panel.

	Serine family			Pyruvate family			Aspartate family					Glutamate family				Other			
Trait	Cys	Gly	Ser	Ala	Leu	Val	Asp	Ile	Lys	Met	Thr	Arg	Glu	Pro	H-pro	His	Phe	TA	PDg
Cys	1.00*	.36*	.48*	.49*	.51*	.41*	.38*	.49*	.26*	.68*	.56*	.44*	.44*	.54*	.07	−.03	.20*	.47*	.12
Gly	.36*	1.00*	.94*	.95*	.93*	.94*	.87*	.93*	.62*	.77*	.93*	.86*	.94*	.42*	.16*	.22*	.46*	.85*	.20*
Ser	.48*	.94*	1.00*	.93*	.94*	.89*	.84*	.93*	.58*	.82*	.95*	.82*	.95*	.47*	.17*	.14	.43*	.85*	.15*
Ala	.49*	.95*	.93*	1.00*	.97*	.96*	.89*	.97*	.64*	.82*	.96*	.89*	.94*	.50*	.15*	.21*	.51*	.88*	.20*
Leu	.51*	.93*	.94*	.97*	1.00*	.95*	.87*	.99*	.64*	.81*	.96*	.88*	.94*	.52*	.12	.19*	.55*	.87*	.16*
Val	.41*	.94*	.89*	.96*	.95*	1.00*	.89*	.96*	.64*	.74*	.91*	.88*	.94*	.46*	.16*	.25*	.48*	.86*	.22*
Asp	.38*	.87*	.84*	.89*	.87*	.89*	1.00*	.86*	.60*	.74*	.84*	.87*	.84*	.44*	.20*	.21*	.46*	.85*	.40*
Ile	.49*	.93*	.93*	.97*	.99*	.96*	.86*	1.00*	.67*	.78*	.96*	.87*	.93*	.49*	.11	.20*	.53*	.86*	.15*
Lys	.26*	.62*	.58*	.64*	.64*	.64*	.60*	.67*	1.00*	.44*	.64*	.65*	.57*	.20*	−.09	.37*	.49*	.56*	.07
Met	.68*	.77*	.82*	.82*	.81*	.74*	.74*	.78*	.44*	1.00*	.86*	.71*	.77*	.49*	.15*	.05	.33*	.73*	.19*
Thr	.56*	.93*	.95*	.96*	.96*	.91*	.84*	.96*	.64*	.86*	1.00*	.85*	.92*	.52*	.15*	.02*	.49*	.86*	.15*
Arg	.44*	.86*	.82*	.89*	.88*	.88*	.87*	.87*	.65*	.71*	.85*	1.00*	.85*	.50*	.12	.30*	.50*	.84*	.39*
Glu	.44*	.94*	.95*	.94*	.94*	.94*	.84*	.93*	.57*	.77*	.92*	.85*	1.00*	.43*	.11	.16*	.42*	.83*	.20*
Pro	.54*	.42*	.47*	.50*	.52*	.46*	.44*	.49*	.20*	.49*	.52*	.50*	.43*	1.00*	.66*	.33*	.38*	.77*	.19*
H-Pro	.07	.16*	.17*	.15*	.12	.16*	.20*	.11	−.09	.15*	.15*	.12	.11	.66*	1.00*	.32*	.00	.55*	.16*
His	−.03	.22*	.14	.21*	.19*	.25*	.21*	.20*	.37*	.05	.20*	.30*	.16*	.33*	.32*	1.00*	.40*	.40*	.06
Phe	.20*	.46*	.43*	.51*	.55*	.48*	.46*	.53*	.49*	.33*	.49*	.50*	.42*	.38*	.00	.40*	1.00*	.52*	.00
TA	.47*	.85*	.85*	.88*	.87*	.86	.85*	.86*	.56*	.73*	.86*	.84*	.83*	.77*	.55*	.40*	.52*	1.00*	.28*
PDg	.12	.20*	.15*	.20*	.16*	.22	.40*	.15*	.07	.19*	.15*	.39*	.20*	.19*	.16*	.06	.00	.28*	1.00*

Note: Correlation coefficients between 19 protein quality traits are listed with their statistical significance (* significant at $p < .05$). Shaded boxes show coefficients of strongly correlated traits ($r > .70$). Abbreviations: Ala, alanine; Arg, arginine; Asp, aspartate; Cys, cystine; Glu, glutamate; Gly, glycine; His, histidine; H-Pro, hydroxyproline; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; PDg, protein digestibility; Phe, phenylalanine; Pro, proline; Ser, serine; TA, total amino acid; Thr, threonine; Val, valine.

by-sequencing data (Rajendran et al., 2021) into SNP genotypes. The *L. culinaris* CDC Redberry Genome Assembly v2.0 (Ramsay et al., 2021) was used as a reference genome. SNPs identified in contigs not incorporated into assembled chromosomes were removed from the analysis. Variants were filtered using VCFtools (Danecek et al., 2011) to include only biallelic SNPs (–min-alleles 2 –max-alleles 2) with a 5% minimum minor allele frequency (–maf 0.05) and a maximum of 20% missing genotypes (–max-missing 0.2). Missing genotypes were then imputed using Beagle 5.4 (Browning et al., 2018). Default parameters were used except for effective population size, which was set to ne = 100,000. Genotypes without AA data were removed, and chromosomes were renamed to integers (1–7) using BCFtools. The final VCF file contained 158 genotypes and 22,280 SNPs.

To mitigate batch effects from the AA and PDg analyses, Bayesian random effects (cf. BLUPs) were used instead of means in the GWAS. Parameter estimates for the effect of genotype were calculated using the rstanarm version 2.21.3 package in R (Goodrich et al., 2020) by fitting the following model:

$$Trait \sim (1|Genotype) + (1|Batch),$$

where *Trait* is the observed trait mean and *Genotype* and *Batch* are random effects.

Genome-wide associations were performed using the Genome Association and Prediction Integrated Tool (GAPIT) version 3 package in R (Wang & Zhang, 2021) using default settings. GAPIT's model selection with the Bayesian information criterion feature determined that the kinship matrix sufficiently accounted for population structure and therefore principal components was not included in the analyses. However, a separate GAPIT analysis was performed to calculate principal component eigenvalues for later visualization with the admixture analysis (Figure 1c,d). Seven GAPIT genome-wide association models were evaluated for fit. These included the Generalized Linear Model (GLM), Mixed Linear Model (MLM), Multiple Loci Mixed Model (MLMM), Compressed MLM (CMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). Figures S2–S6 are Quantile–Quantile and Manhattan plots from these models. The MLMM (Segura et al., 2012) demonstrated superior fit across most traits and was selected for further analysis. A Bonferroni threshold ($0.05/22{,}280 = 2.24 \times 10^{-6}$) was used to determine significance. Manhattan plots (Figures 2, S5, and S6) and QQ plots (Figures S2–S4) were drawn using the CMplot version 4.1.0 package in R (https://github.com/YinLiLin/CMplot).

LD blocks were determined using PLINK v1.07 (Purcell et al., 2007) by calculating pair-wise correlations (r^2) of significant SNPs with adjacent SNPs within a 1-Mb window. LD blocks were determined to decay either at the first SNP with $r^2 < 0.4$ or at 100 kb past the final linked SNP, whichever was less. The 100-kb provision was to help account for the low SNP density of many regions of the genome that otherwise resulted in highly inflated LD block sizes. Genes within local LD with significant SNPs were identified using a custom python script (https://github.com/jlboat/features_from_snps) and were considered candidate genes.

2.6 | Population structure and origin analysis

Population structure was estimated using ADMIXTURE (Alexander & Lange, 2011). The optimal number of ancestral populations ($K = 6$) was determined by selecting the model with the lowest cross-validation error using fivefold cross-validation. The model generated a Q matrix containing ancestral coefficients for each genotype. Accessions were categorized into subpopulations based on their highest ancestry coefficient (> 0.5). An admixture plot (Figure 1b) was drawn using the R package gglot2 version 3.3.6 (Wickham & Wickham, 2016).

The global distribution of accessions by ancestral subpopulation was then visualized (Figure 1a). ISO3 country codes were first assigned to accessions resulting in 52 unique countries of origin. Origin information was missing from two accessions resulting in the inclusion of 156 accessions in the figure. The ddply function in the R package plyr version 1.8.7 (Wickham, 2011) was used to calculate the mean ancestral coefficients for each country of origin. These values were then multiplied by the number of accessions per country before being translated into a SpatialPolygonsDataFrame using the function joinCountryData2Map in rworldmap version 1.3–4 (South, 2011). The mapPies function in rworldmap was then used to draw a figure displaying ancestral population pie charts for each country of origin (Figure 1a). These pie charts depict the average admixture composition of accessions from the same country of origin. Pie chart circumferences are proportional to the number of accessions sharing a country of origin. Principal component analysis (PCA) scatter plots with accessions classified by ADMIXTURE subpopulation were drawn in ggplot2 using PCA variance components calculated in GAPIT (Figure 1c,d).

Analysis of variance was performed to determine if ancestral group had a significant effect on protein quality traits. Trait means across accession replicates, previously calculated in JMP, were combined with ADMIXTURE subpopulations, and a model was developed in JMP with trait concentration as the response variable and subpopulation as a fixed effect. For models with a significant effect (Ala, Arg, Cys, and His), Fisher's protected least squares difference (LSD) procedure was used to determine differences between subpopulations. Boxplots were drawn in JMP (Figure 3). All figures received final formatting using Adobe Illustrator 2019.

3 | RESULTS

3.1 | Summary statistics and correlations

AA concentration means ranged from 0.21 to 4.45% (Table 1). S-AAs, Met and Cys, had mean concentrations of 0.21 and 0.22%, respectively. The four AAs with the highest mean concentrations were Arg (2.72%),

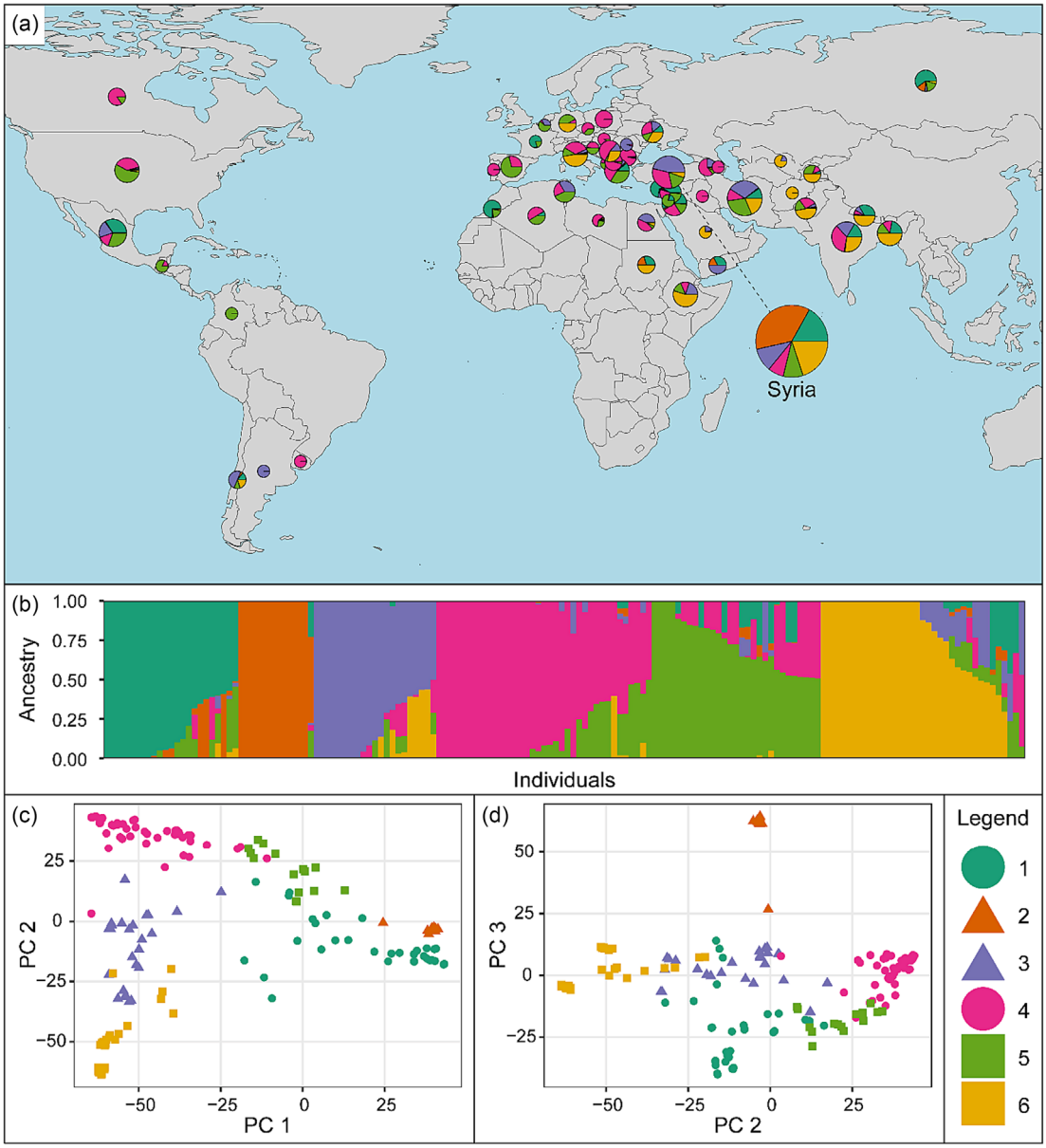

FIGURE 1 Lentil population origin and population structure analysis based on 22,280 single nucleotide polymorphisms (SNPs) used in the genome-wide association study. (a) Pie charts depict average admixture composition of accessions from the same country of origin. Pie chart circumferences are proportional to the number of accessions sharing each country of origin. The colors depict the average ancestral subpopulation composition of each location as determined by ADMIXTURE analysis where $k = 6$. (b) The bar chart shows the admixture of accessions, which are sorted into by their ancestral subpopulation classification. (c–d) Scatter plots depicting the first three principal components with points representing accessions that have been colored corresponding to their ADMIXTURE ancestral subpopulation classification.

H-Pro (3.16%), Asp (3.87%), and Glu (4.45%). The mean total AA concentration was 29.0%. The mean PDg was 88% (Table 1). Figure S1 displays histograms representing protein quality trait distributions as they compare to normal density curves of the same mean and standard deviation. However, the data distributions for Cys and, to a lesser degree, Ile and Lys appear bimodal. As estimated from trait means, a 100 g serving of lentil would provide 100% of the RDA of Ile, Leu, Lys, Phe, and Thr (Table 1). The same serving would provide 25% and 43% of the RDA of Val and TA, respectively. The lowest %RDA values were for lentil's limiting AAs, Met, and Cys, at 19 and 20%, respectively; these are the S-AAs. Repeatability estimates for protein quality traits ranged from 3.8% to 34.9% (Table 1). The seven highest estimates were PDg (34.9%), Met (26.4%), TA (23.5%), Thr (22.4%), Asp (22.2%), Ile (21.5%), and Leu (21.4%). Repeatability estimates for ratios of individual AAs to TA ranged from 0% to 27% (Table S3). These estimates were 4.8% lower on average than their respective non-ratio AAs. Met: TA was the exception, with a slightly higher estimate (27%) than Met (26.4%). The following accessions are best for incorporation into the breeding programs to increase AA concentrations with higher PDg—ILL 7727, ILL 858, ILL 1097, and ILL 595 (Table S4).

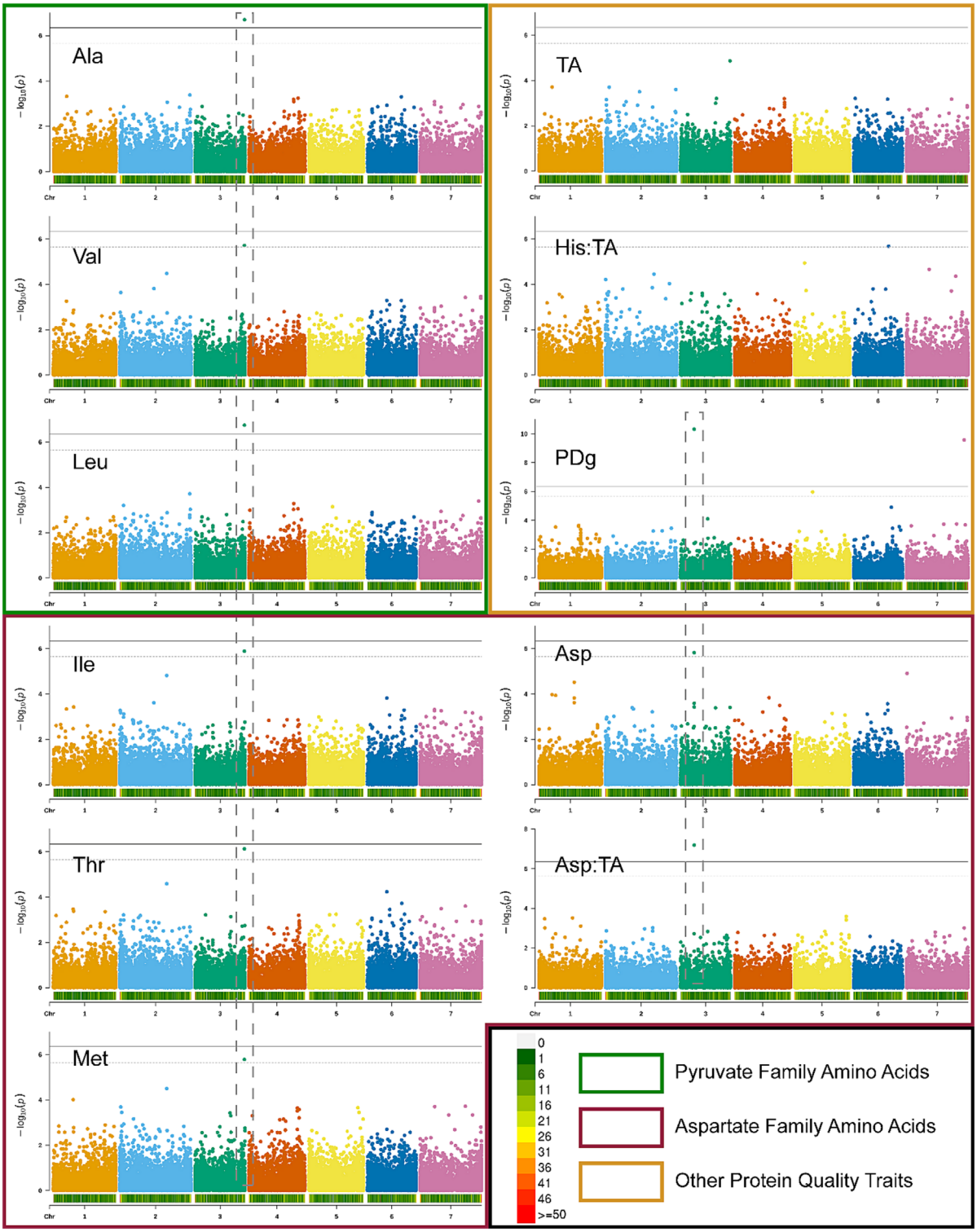

FIGURE 2 Manhattan plots depicting genome-wide association study single-nucleotide (SNP) association results. Chromosome numbers are across the *x*-axis with the −log(*p*-value) on the *y*-axis. Points represent SNPs associated with each trait and are colored according to the chromosome on which they reside. SNPs exceeding the dotted and solid gray horizontal lines were significantly associated with a trait at $p < 0.05/22{,}280$ and $p < 0.01/22{,}280$, respectively (Bonferroni correction). Gray dashed line boxes indicate loci that shared significant association across multiple traits. Colored outlines represent pyruvate family amino acids, aspartate family amino acids, and other protein quality traits (Gly, His:TA, PDg). SNP density plots are located above chromosome numbers on a red to green scale of 1 to 50 SNPs per 1 Mb. Ala, alanine; Asp, aspartate; Ile, isoleucine; Leu, leucine; Met, methionine; Thr, threonine; Val, valine; TA, total amino acid; PDg, protein digestibility; His:TA, histidine to total amino acid; Asp:TA, aspartate to total amino acid.

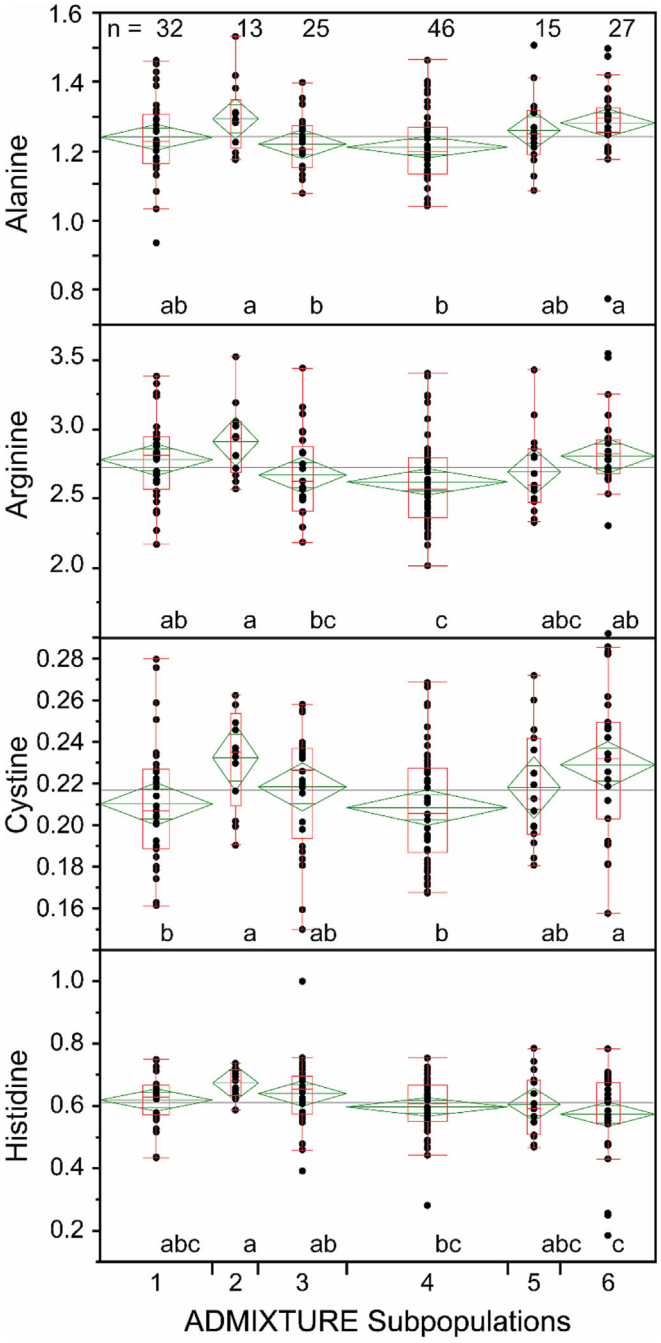

FIGURE 3 Ancestral subpopulations from ADMIXTURE analysis were compared based on four amino acid concentrations using one-way analysis of variance and Fisher's protected least squares difference (LSD). Boxplots connected by different letters have significantly different means ($p < .05$). The solid gray horizontal line represents the overall mean of the trait. Green diamonds indicate the 95% confidence interval of the mean for each subpopulation. The diamond width is proportional to the number of samples belonging to the subpopulation ($n_1 = 32$, $n_2 = 13$, $n_3 = 25$, $n_4 = 46$, $n_5 = 15$, $n_6 = 27$).

The following traits were strongly correlated with one another ($r > .7$): Gly, Ser, Ala, Leu, Val, Asp, Ile, Met, Thr, Arg, Glu, and TA (Table 2). Lys had a moderate correlation ($r = .56$–$.67$) with most of the strongly correlated traits, except a low correlation with Met ($r = .44$). Leu and Ile had the strongest correlation at .99. The S-AAs, Cys and Met, had a moderate correlation of .68. Pro and H-Pro also had a moderate correlation of .66. PDg had only three correlations with $r > .23$; these were Asp ($r = .40$), Arg ($r = .39$), and TA ($r = .28$). Most trait correlations were significant at $p < .05$; however, H-Pro, His, and PDg were noteworthy for having seven, four, and four insignificant correlations, respectively.

3.2 | Population structure and subpopulation trait differences

The lentil diversity panel was determined to have six ancestral subpopulations by ADMIXTURE analysis (Figure 1b). Subpopulations 1 through 6 were composed of 32, 13, 25, 46, 15, and 27 accessions, respectively. Mostly subtle visual associations were seen between subpopulations and regions of origin (Figure 1a). Subpopulation six was mostly absent from countries in North and South America. Accessions from the United States and Canada were composed mostly of clusters four and five, whereas Syrian accessions saw substantial representation of all six subpopulation clusters. Subpopulation two showed little admixture with the other subpopulations (Figure 1b) and was almost exclusively represented by the accessions originating from Syria (Figure 1a), which is the center for the ICARDA lentil breeding program. The first three principal components (PCs) of the PCA accounted for 13.9, 10.1, and 5.1% of the total variance. Clear separation of ADMIXTURE clusters was seen in the PCA scatter plots (Figure 1c,d). Accessions within subpopulation two were tightly clustered on the PCA scatterplots and were clearly delineated from the other clusters by PC3 (Figure 1d).

When analysis of variance was performed between ADMIXTURE clusters for each protein quality trait, significantly different ($p < .05$) means were identified for Ala, Arg, Cys, and His (Figure 3). Pair-wise comparison showed that cluster two had a mean in the highest letter category across all four traits, whereas cluster four had a mean in the lowest letter category across all four traits. Thus, across these four traits, clusters two and four always had significantly different means. Ala clusters two and six had significantly higher means than clusters three and four. Arg clusters one, two, and six had significantly higher means than cluster four, whereas clusters three and four had significantly lower means than cluster two. Cys clusters two and six had significantly higher means than clusters one and four. His clusters two and three had significantly higher means than cluster six, whereas clusters four and six had significantly lower means than cluster two.

3.3 | Genome-wide association studies

Five significantly associated lead SNPs (LD blocks) were identified across 10 protein quality traits (Table 3 and Figure 2). These SNPs had post-filtering minor allele frequencies below 0.05 and were in local LD with 27 genes (Tables 3 and S4). QQ plots showed the MLMM to

TABLE 3 Lead SNPs associated with protein quality traits from the lentil diversity panel.

Lead SNP[a]	LD block size (kb)	Traits	R^2 (%) [b]	Candidate genes[c]	Gene descriptions[d]
SLcu.2RBY.Chr3_115494955	818	Aspartate,	13.1	Lcu.2RBY.3g018060	Uncharacterized protein
		Aspartate: TotalAA,	17.1		
		Protein digestibility	0.0	Lcu.2RBY.3g018070	Glutathione S-transferase
				Lcu.2RBY.3g018080	Glutathione S-transferase
				Lcu.2RBY.3g018090	Glutathione S-transferase
				Lcu.2RBY.3g018100	Glutathione S-transferase; amino-terminal domain protein
				Lcu.2RBY.3g018110	Uncharacterized protein
				Lcu.2RBY.3g018120	Uncharacterized protein
				Lcu.2RBY.3g018130	Eukaryotic aspartyl protease family protein
				Lcu.2RBY.3g018140	60S ribosomal protein L18a
				Lcu.2RBY.3g018150	3-hydroxyisobutyryl-CoA hydrolase-like protein
				Lcu.2RBY.3g018160	Polyprotein
				Lcu.2RBY.3g018170	Subtilisin-like serine protease
				Lcu.2RBY.3g018180	Ulp1 protease family, carboxy-terminal domain protein
				Lcu.2RBY.3g018190	Uncharacterized protein
				Lcu.2RBY.3g018200	Lipid transfer protein
SLcu.2RBY.Chr3_424796277	117	Alanine,	15.9	Lcu.2RBY.3g073770	Gibberellin 2-beta-dioxygenase
		Isoleucine,	14.3		
		Leucine,	16.2	Lcu.2RBY.3 g073780	Gibberellin 2-beta-dioxygenase
		Methionine,	13.3	Lcu.2RBY.3g073790	Stem 28-kDa glycoprotein
		Threonine,	14.3	Lcu.2RBY.3g073800	Receptor-like kinase plant
		Valine	13.7		
SLcu.2RBY.Chr5_155570229	100	Protein digestibility	0.7	Lcu.2RBY.5g028750	Alpha/beta-hydrolase superfamily protein
				Lcu.2RBY.5g028760	Uncharacterized protein
SLcu.2RBY.Chr6_301590681	0	Histidine: TotalAA	15.4	Lcu.2RBY.6g042780	Alpha-mannosidase
				Lcu.2RBY.6g042790	RNA-directed DNA polymerase (reverse transcriptase) chromo zinc finger, CCHC-type peptidase aspartic, active site Polynucleotidyl transferase, ribonuclease H fold
SLcu.2RBY.Chr7_497443978	69	Protein digestibility	0.2	Lcu.2RBY.7g065930	Transmembrane protein
				Lcu.2RBY.7g065940	Hypothetical protein
				Lcu.2RBY.7g065950	Peptide/nitrate transporter
				Lcu.2RBY.7g065960	Glycerophosphoryl diester phosphodiesterase family protein

Note: One or more traits were associated with each locus. Ratios of an amino acid with the total amino acid concentration are indicated with a colon (e.g., Aspartate:TotalAA). The loci are identified by the lead single nucleotide polymorphism (SNP) of its corresponding linkage disequilibrium (LD) block. The genes within each block are listed along with their descriptions.

[a]The SNP at a locus exceeding a significance threshold of $p < .05/22{,}280$ (Bonferroni correction) in association with a trait.

[b]Difference between R^2 of GAPIT GLM model with and without lead SNP.

[c]Genes in local LD with significant SNPs.

[d]See supporting information data (Table S2) for gene annotation source information.

control well for population structure and false positives for most traits compared to other models (Figures S2–S4). Two lead SNPs were associated with multiple traits (Table 3 and Figure 2, gray dashed boxes). The first of these SNPs, SLcu.2RBY.Chr3_115394955, was associated with PDg ($R^2 < 0.1\%$) and two aspartate family traits (Asp [$R^2 = 13.1\%$] and Asp:TA [$R^2 = 17.1\%$]). The LD block containing the SNP was 818 kb in size and contained 15 genes including four glutathione S-transferase genes and three protease family protein genes (Table 3). The second SNP, SLcu.2RBY.Chr3_424696277, was associated with three pyruvate family AAs (Ala [$R^2 = 15.9\%$], Val

[$R^2 = 13.7\%$], and Leu [$R^2 = 16.2\%$]) and three aspartate family AAs (Ile [$R^2 = 14.3\%$], Thr [$R^2 = 14.3\%$], and Met [$R^2 = 13.3\%$]) (Table 3 and Figure 2). The LD block containing the SNP was 117 kb in size and contained four genes: Lcu.2RBY.3 g073770 (gibberellin-2-beta-dioxygenase), Lcu.2RBY.3 g073780 (gibberellin-2-beta-dioxygenase), Lcu.2RBY.3 g073790 (stem 28-kDa glycoprotein), and Lcu.2RBY.3 g073800 (plant receptor-like kinase). SNP densities were visualized by the green to red scaled plots above chromosome numbers in Figures 2, S5, and S6. Densities were relatively low (0–21 SNPs per Mb) across most of the genome, with regions near the end of chromosomes showing higher SNP densities (>30 SNPs per Mb).

4 | DISCUSSION

AA concentrations agreed well with the U.S. Department of Agriculture's reference values (U.S. Department of Agriculture, 2019). The reference mean fell within the ranges reported here, with the exceptions of Cys (0.32% [USDA] vs. 0.15–0.29% [Table 1]) and Val (1.22% [USDA] vs. 0.14–0.26% [Table 1]), where the reference was higher, and Pro (1.03% [USDA] vs. 1.44–3.96% [Table 1]), where the reference was lower. Given that a reference value is commonly based on only one or a few varieties, the values for these amino acids may simply not be representative of other lentil accessions, especially given the wide variation in the literature. However, there are known analysis limitations that may contribute to the discrepancy. Cys is an S-AA known to degrade during acid hydrolysis. The method used included a pre-oxidative step using performic acid. This was intended to convert all Cys to cysteic acid, which is a more stable derivative. Nonetheless, some Cys is expected to degrade, which may be the case here. Val has long been noted as being resistant to acid digestion when in a Val–Val or Val–Ile peptide bond, so this may help explain the low Val concentration (Nair et al., 1976). Although Pro is high compared to the reference, this concentration agrees with the literature (Salaria et al., 2022). Notably, many of the standard reference values are low compared to the literature. Means and ranges also agreed well with values estimated by near-infrared spectroscopy, with the exceptions of high Arg and Pro and low Lys and Val (Hang, 2021). The mean TA concentration (29%; Table 1) agreed well with ranges found in the literature for protein content (U.S. Department of Agriculture, 2019). The PDg estimate (88%, Table 1) accords well with the range of 50–95% found in the literature (Martín-Cabrejas et al., 2009; Monsoor & Yusuf, 2002; Shekib et al., 1986). Percent RDA values reinforce that lentil is a good source of protein and essential AAs, except for the limiting AAs, Met and Cys.

Repeatability is considered the upper bound of broad-sense heritability (Kruijer et al., 2014). Protein content in food legumes is significantly affected by environment and genotype × environment effects (Pratap & Kumar, 2011). Heritability estimates (broad sense) of AA and protein concentrations vary across legume species and studies. The TA repeatability estimate presented here (23.5%, Table 1) is low within this range. Heritability estimates of individual AAs are not widely reported in legumes. However, in soybean, heritability estimates of individual AAs ranged from 40.9% (Trp) to 81.8% (Asp) (Jiang & Katuuramu, 2021), whereas Met was high (99.7%) in chickpea (Desai et al., 2015). The large difference in heritability estimates between these studies and the present study is likely because of differences in study design. These studies used field plots instead of potted plants and included 54 and 48 genotypes, respectively, compared to 183 genotypes here. Repeatability estimates (Table 1) ranged from 3.8% (Phe) to 26.4% (Met). Lentil showed low heritability for Phe compared to the other AAs, and the values are lower than the reported for soybean; further studies are required to understand the AA differences between soybean and lentil. One speculation may be that soybean is a warm season, which has a different nitrogen and AA biochemistry than the cool season food legumes like lentil. The heritability of protein content in food legumes is moderate to high, ranging from 20% to 85% (Baudoin & Maquet, 1999; Patil et al., 2020; Pratap & Kumar, 2011). Estimates in lentil fall within this range (Gautam et al., 2018). The PDg repeatability estimate was 34.9% (Table 1), which was the highest of the protein quality traits. PDg heritability estimates are not widely reported for legumes. However, heritability estimates in sorghum range from 91% to 96% (Abdelhalim et al., 2019; Pfeiffer, 2017). Repeatability estimates for lentil protein quality traits were low to moderate and comparable to literature values. This indicates that some of these traits, such as Met and PDg, are breeding target traits for increased protein quality but need more testing in the field to understand the heritability of these traits. A recent study reported that dry pea (*Pisum Sativum* L.) cultivars grown under organic field conditions showed 11.8 to 22.2 g/100 g for total AAs (liberated), 12.6 to 27.6 g/100 g for total protein, and 83 to 95% for in vitro protein digestibility. Dry pea contained a range of individual AAs, including nine essential AAs, with a mean of 0.22 g/100 g for SAAs and 0.88 g/100 g for Lys (Thavarajah et al., 2023). These organic dry pea cultivars provide a significant amount of the recommended daily allowance (%RDA) of several AAs (14–66% His, 79–138% Iso, 76–169% Leu, 57–147% Lys, 15–85% Met + Cys, 76–189% Phe + Tyr, and 94–169% Val) as well as protein (22–48%; Thavarajah et al., 2023). Overall, lentil and dry pea have similar protein quality and significant amounts of essential AAs and thus have good potential for future plant-based food production.

Strong correlations were seen between several protein quality traits, which is consistent with the literature (Hang, 2021; Wang & Daun, 2006). The present study found strong correlations ($r > 0.7$) among Gly, Ser, Ala, Leu, Val, Asp, Ile, Met, Thr, Arg, Glu, and TA concentration (Table 2). Cys had low to moderate correlations ($r = .2$–$.7$) with other AAs here and as noted by Hang (2021). Interestingly, the trait that correlated strongest with Cys was Met, the other S-AA. PDg was only weakly correlated ($r < .39$) with other protein quality traits. However, its strongest correlation was with Arg ($r = .39$), which is one of the AAs incorporated into the PDg calculation. As will be discussed below, several strongly correlated AAs share significantly associated SNPs (Table 3).

ADMIXTURE analysis that determined the optimal number of ancestral subpopulations was $k = 6$ (Figure 1b). This is comparable but higher than other admixture analyses using the software

STRUCTURE, which found between three and five ancestral subpopulations for *L. culinaris* Medik. (Khazaei et al., 2016; Kumar et al., 2019; Liber et al., 2021; Pavan et al., 2019; Rajendran et al., 2021). The absence of subpopulation six from most of the accessions in North and South America suggests either that germplasm from this subpopulation was not widely incorporated into these regions or that North and South American accessions from subpopulation six were simply not included in the present study. North and South American accessions were primarily from subpopulations four and five. In contrast, Syrian accessions (the largest number at $n = 35$) contain substantial representation of all six subpopulations. This is not surprising because ICARDA, the source of the populations used in this study, is located in Syria. Subpopulation two is distinct because it shows relatively little admixture compared to the other subpopulations (Figure 1b), which suggests a highly related (inbred) group of accessions that is relatively distant genetically from the other subpopulations. This is further confirmed by the PCA (Figure 1c,d), which shows that accessions classified as subpopulation two are tightly clustered and clearly delineated from the other clusters by PC 3 (Figure 1d). Interestingly, subpopulation two is also primarily represented in the accessions originating from Syria.

Analysis of variance showed significant differences between the means of subpopulations across accessions for Ala, Arg, Cys, and His (Figure 3). Subpopulations two and four had the highest and lowest respective means for each trait. This suggests that genetically and phenotypically divergent accessions could be selected from these subpopulations for recombinant population development. Additionally, high Ala and Arg accessions could be selected from subpopulations one, two, five, or six; high Cys accessions could be selected from subpopulations two, three, five, or six; and high His accessions could be selected from subpopulations one, two, three, or five.

GWAS resulted in the identification of five significantly associated leading SNPs in local LD with 27 genes across 10 protein quality traits (Table 3 and S4). The significant SNPs had minor allele frequencies below 0.05 after filtering (Table S4). The number of false positives in GWAS increases at lower minor allele frequencies (Tabangin et al., 2009); however, many causative mutations are expected to occur at low frequencies in a population because of purifying selection (Tibbs Cortes et al., 2021). Consequently, these SNPs and the genes in local LD with them should be investigated for their effect on protein quality traits but with informed caution. LD blocks contained only one or two SNPs per block. This is not surprising because of the low SNP density observed here (22,280 SNPs/3.69-Gb reference genome = ~ 6 SNPs/Mb; Figure 2). However, this could also be indicative of false positives.

Two leading SNPs are noteworthy because they were associated with multiple traits (Table 3 and Figure 2). Furthermore, a relatively high R^2 difference was found between the GWAS model with and without the lead SNP for most traits ($R^2 = 13.1$ to 17.1%), suggesting that these markers may be useful for marker-assisted selection after further evaluation. PDg was an exception to this ($R^2 < 0.1\%$) The first SNP, SLcu.2RBY.Chr3_115394955, was associated with Asp, Asp:TA, and PDg. Interestingly, Asp was only weakly correlated with PDg ($r = 0.40$, Table 2). Although individual AAs and their corresponding ratio with TA (such as Asp and Asp:TA) might be expected to share significant associations, this was the only pair that was identified. (It might also be expected that ratios of TA may share associations with TA; however, because TA was not significantly associated with any marker, this was not the case.) The LD block containing SLcu.2RBY.Chr3_115394955 was 818 kb in size and contained 15 genes. Four genes were identified as glutathione S-transferase genes by homology. Glutathione S-transferase genes are a supergene family whose products aid in neutralizing toxins by helping facilitate the anti-oxidative activity of glutathione (Gullner et al., 2018). These genes are upregulated during stress, as has been shown in lentil under arsenic stress (Talukdar, 2016). A glutathione S-transferase gene has also been proposed as a candidate gene for Verticillium wilt disease resistance in *Arabidopsis* (Gong et al., 2018). The role of glutathione S-transferase in protein quality remains unclear. SLcu.2RBY.Chr3_115394955 was also in local LD with three protease family genes. Proteases degrade unwanted proteins and maintain protein quality in plant cells (García-Lorenzo et al., 2006). Protease inhibitors in grain reduce the activity of protease enzymes during animal digestion; consequently, seed with higher concentrations of protease inhibitors has lower PDg (Singh & Jambunathan, 1981). Regulation of protease genes and protease inhibitor genes is hypothesized to be synchronized within the plant, leading to this association with PDg.

The second leading SNP associated with multiple traits was SLcu.2RBY.Chr3_424696277, which was associated with Ala, Ile, Leu, Met, Thr, and Val (Figure 2). These traits were strongly correlated (Table 2). The block contained four genes, two of which were gibberellin 2-beta-dioxygenase genes. This gene family is involved in numerous developmental processes in plants, such as seed germination, leaf expansion, shoot/stem lengthening, and reproductive structures and processes (Wang et al., 2014). Therefore, these genes would likely affect protein quality traits; however, altering expression may affect numerous traits besides protein quality traits. The other two genes were a stem 28-kDa glycoprotein gene and a plant receptor-like kinase, both of which are broad descriptors requiring functional analysis for further investigation. Many of the associated genes across all traits were either for uncharacterized or poorly characterized proteins. Others, such as glutathione S-transferase, were well characterized as a protein family but not for the specific function of the protein in biochemical pathways. This underscores the ongoing need for functional characterization of plant proteins and the relative lack of annotation specific to lentil.

This current study was conducted in the greenhouse for a single growing season to understand the genetic diversity of protein quality of these accessions under controlled conditions. Therefore, these results should be considered preliminary because of the quantitative nature of these traits, which need to be tested over several growing locations and years. Future studies will confirm genomic associations across environments because seed protein quality traits may be influenced by environmental and G × E effects. To this end, additional replication in multiple environments should be pursued. Additionally, functional analysis of candidate genes, especially glutathione

S-transferase, protease family, and gibberellin 2-beta-dioxygenase genes are needed. Once accomplished, breeding markers can be identified for these genes for marker-assisted selection. Other phenotyping methods, such as Fourier-Transform Infrared Spectroscopy (FTIR), should be pursued to lower costs and increase sampling throughput, which would make evaluating larger population sizes and increased replication more tenable.

Globally, animal proteins become more expensive because of economic and environmental costs; therefore, plant-based protein sources, such as lentil, will become increasingly important to global food systems to provide essential AAs and micronutrients at a low cost. Biofortification of lentils for increased protein quality and digestibility is a sustainable approach to providing this highly nutritious food at a low cost for vulnerable populations around the world. Our study results clearly indicated that lentil contains ~19–36% TA with 86–91% PDg that provide significant amounts of %RDA for essential AAs. Consequently, evaluating and targeting protein quality traits for genomic-assisted lentil breeding is a valuable objective for lentil breeding programs to develop protein-biofortified cultivars to combat protein malnutrition.

AUTHOR CONTRIBUTIONS

Nathan Johnson: doctoral student worked on the project, experimental design, analytical chemistry, bioinformatics, wrote and edited the manuscript; **J Lucas Boatwright**: bioinformatics and manuscript edits; **William Bridges**: statistical analysis and manuscript edits; **Pushparajah Thavarajah**, analytical chemistry and manuscript edits; **Shiv Kumar:** experimental design, assemble the diversity panel and manuscript edits; **Dil Thavarajah**: project director, conceptualization, doctoral mentor for Nathan Johnson and wrote/edited the manuscript with Nathan Johnson.

ACKNOWLEDGMENTS

Funding support for this project was provided by the Organic Agriculture Research and Extension Initiative (OREI) (award no. 2018-51300-28431/proposal no. 2018-02799) and the United States Department of Agriculture, National Institute of Food and Agriculture (award no. /proposal no. 2021-02927) (DT); the USDA National Institute of Food and Agriculture, [Hatch] project [1022664] (DT); the Good Food Institute (DT); SC Department of Agriculture; and FoodShot Global. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the USDA.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

Code, data, and supplemental files can be found at https://github.com/njohns4/LentilProteinQualityGWAS. Supporting information tables and figures, including lentil population information, protein quality trait data, and full GAPIT MLMM GWAS results (Tables S2–S4), are also available at *G3* online. The *Lens culinaris* CDC Redberry Genome Assembly v2.0 is available at https://knowpulse.usask.ca/genome-assembly/Lcu.2RBY.

ORCID

Nathan Johnson https://orcid.org/0000-0002-4537-7966
J. Lucas Boatwright https://orcid.org/0000-0003-4371-2457
Shiv Kumar https://orcid.org/0000-0001-8407-3562
Dil Thavarajah https://orcid.org/0000-0002-4251-7476

REFERENCES

Abdelhalim, T. S., Kamal, N. M., & Hassan, A. B. (2019). Nutritional potential of wild sorghum: Grain quality of Sudanese wild sorghum genotypes (Sorghum bicolor L. Moench). *Food Science & Nutrition*, *7*, 1529–1539. https://doi.org/10.1002/fsn3.1002

Agilent Application Note. (2010). *Separation of two sulfurated amino acids with other seventeen amino acids by HPLC with pre-column derivatization*. Agilent Technologies.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, 246. https://doi.org/10.1186/1471-2105-12-246

Alghamdi, S. S., Khan, A. M., Ammar, M. H., El-Harty, E. H., Migdadi, H. M., El-Khalik, S. M. A., Al-Shameri, A. M., Javed, M. M., & Al-Faifi, S. A., S. A. (2014). Phenological, nutritional and molecular diversity assessment among 35 introduced lentil (Lens culinaris Medik.) genotypes grown in Saudi Arabia. *International Journal of Molecular Sciences*, *15*, 277–295. https://doi.org/10.3390/ijms15010277

Bailey-Serres, J., Fukao, T., Ronald, P., Ismail, A., Heuer, S., & Mackill, D. (2010). Submergence tolerant rice: SUB1's journey from landrace to modern cultivar. *Rice*, *3*, 138–147. https://doi.org/10.1007/s12284-010-9048-5

Barbosa, P., Melnyk, S., Bennuri, S. C., Delhey, L., Reis, A., Moura, G. R., Børsheim, E., Rose, S., & Carvalho, E. (2021). Redox imbalance and methylation disturbances in early childhood obesity. *Oxidative Medicine and Cellular Longevity*, *2021*, 1–16, 2207125. https://doi.org/10.1155/2021/2207125

Baudoin, J. P., & Maquet, A. (1999). Improvement of protein and amino acid contents in seeds of food legumes. A case study in Phaseolus. *Biotechnology, Agronomy, Society and Environment*, *3*, 220–224.

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *American Journal of Human Genetics*, *103*, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., Wang, S., Shi, L., Zhou, B., Li, Z., Peng, X., Yang, C., Qu, L., Liu, X., & Luo, J. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nature Communications*, *7*, 12767. https://doi.org/10.1038/ncomms12767

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., & Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: Lessons from the breeder's equation. *Theoretical and Applied Genetics*, *132*, 627–645. https://doi.org/10.1007/s00122-019-03317-0

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Desai, K., Tank, C., Gami, R., & Patel, A. (2015). Genetic variability in indigenous collection of chickpeas (Cicer arietinum L.) genotypes for seed yield and quality traits. *Journal of Progressive Agriculture*, *9*, 59–62.

Diniz, A. L., Giordani, W., Costa, Z. P., Margarido, G. R. A., Perseguini, J. M. K. C., Benchimol-Reis, L. L., Chiorato, A. F., Garcia, A. A. F., & Vieira, M. L. C. (2019). Evidence for strong kinship influence on the extent of linkage disequilibrium in cultivated common beans. *Genes (Basel)*, *10*(1), 5. https://doi.org/10.3390/genes10010005

Dissanayake, R., Cogan, N. O. I., Smith, K. F., & Kaur, S. (2021). Application of genomics to understand salt tolerance in lentil. *Genes (Basel)*, *12*, 1–18. https://doi.org/10.3390/genes12030332

Evans, J. R., & Lawson, T. (2020). From green to gold: Agricultural revolution for food security. *Journal of Experimental Botany*, *71*, 2211–2215. https://doi.org/10.1093/jxb/eraa110

Foyer, C. H., Lam, H. M., Nguyen, H. T., Siddique, K. H. M., Varshney, R. K., Colmer, T. D., Cowling, W., Bramley, H., Mori, T. A., Hodgson, J. M., Cooper, J. W., Miller, A. J., Kunert, K., Vorster, J., Cullis, C., Ozga, J. A., Wahlqvist, M. L., Liang, Y., Shou, H., … Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants*, *2*, 1–10, 16112. https://doi.org/10.1038/NPLANTS.2016.112

García-Lorenzo, M., Sjödin, A., Jansson, S., & Funk, C. (2006). Protease gene families in Populus and Arabidopsis. *BMC Plant Biology*, *6*, 1–24, 30. https://doi.org/10.1186/1471-2229-6-30

Gautam, N. K., Bhardwaj, R., Yadav, S., Suneja, P., Tripathi, K., & Ram, B. (2018). Identification of lentil (Lens culinaris Medik.) germplasm rich in protein and amino acids for utilization in crop improvement. *Indian Journal of Genetics and Plant Breeding*, *78*, 470–477. https://doi.org/10.31742/IJGPB.78.4.9

Gehrke, C. W., Wall, L. L. Sr., Absheer, J. S., Kaiser, F. E., & Zumwalt, R. W. (1985). Sample preparation for chromatography of amino acids: Acid hydrolysis of proteins. *Journal of the Association of Official Analytical Chemists*, *68*, 811–821. https://doi.org/10.1093/jaoac/68.5.811

Ghumman, A., Singh, N., Kaur, A., & Rana, J. C. (2019). Diversity in protein secondary structure, molecular weight, mineral and amino acid composition of lentil and horse gram germplasm. *Journal of Food Science and Technology*, *56*, 1601–1612. https://doi.org/10.1007/s13197-019-03676-y

Gilbert, J. A., Bendsen, N. T., Tremblay, A., & Astrup, A. (2011). Effect of proteins from different sources on body composition. *Nutrition, Metabolism, and Cardiovascular Diseases*, *21*, B16–B31. https://doi.org/10.1016/j.numecd.2010.12.008

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*, e90346. https://doi.org/10.1371/journal.pone.0090346

Gong, Q., Yang, Z., Chen, E., Sun, G., He, S., Butt, H. I., Zhang, C., Zhang, X., Yang, Z., du, X., & Li, F. (2018). A phi-class glutathione S-transferase gene for Verticillium wilt resistance in Gossypium arboreum identified in a genome-wide association study. *Plant & Cell Physiology*, *59*, 275–289. https://doi.org/10.1093/pcp/pcx180

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. *R package version 2.21.1*. https://mc-stan.org/rstanarm/

Gullner, G., Komives, T., Király, L., & Schröder, P. (2018). Glutathione S-transferase enzymes in plant-pathogen interactions. *Frontiers in Plant Science*, *871*, 1–19, 1836. https://doi.org/10.3389/fpls.2018.01836

Hang, J. (2021). Genome-wide association study of seed protein and amino acid contents in cultivated lentils as determined by near-infrared reflectance spectroscopy [Master's Thesis, University of Manitoba, Canada]. mspace.lib.umanitoba.ca

Hang, J., Shi, D., Neufeld, J., Bett, K. E., & House, J. D. (2022). Prediction of protein and amino acid contents in whole and ground lentils using near-infrared reflectance spectroscopy. *LWT*, *165*, 113669. https://doi.org/10.1016/j.lwt.2022.113669

Jalgaonkar, S., Gajbhiye, S., Sayyed, M., Tripathi, R., Khatri, N., Parmar, U., & Shankar, A. (2022). S-adenosyl methionine improves motor co-ordination with reduced oxidative stress, dopaminergic neuronal loss, and DNA methylation in the brain striatum of 6-hydroxydopamine-induced neurodegeneration in rats. *Anatomical Record*, *1–11*, 820–830. https://doi.org/10.1002/ar.24948

Jiang, G. L., & Katuuramu, D. N. (2021). Comparison of seed fatty and amino acids in edamame dried using two oven-drying methods and mature soybeans. *Journal of the Science of Food and Agriculture*, *101*, 1515–1522. https://doi.org/10.1002/jsfa.10766

Johnson, N., Boatwright, J. L., Bridges, W., Thavarajah, P., Kumar, S., Shipe, E., & Thavarajah, D. (2021). Genome-wide association mapping of lentil (Lens culinaris Medikus) prebiotic carbohydrates toward improved human health and crop stress tolerance. *Scientific Reports*, *1–12*, 13926. https://doi.org/10.1038/s41598-021-93475-3

Johnson, N., Johnson, C. R., Thavarajah, P., Kumar, S., & Thavarajah, D. (2020). The roles and potential of lentil prebiotic carbohydrates in human and plant health. *Plants, People, Planet*, *00*, 1–10. https://doi.org/10.1002/ppp3.10103

Karaca, N., Ates, D., Nemli, S., Ozkuru, E., Yilmaz, H., Yagmur, B., Kartal, C., Tosun, M., Ozdestan, O., Otles, S., Kahriman, A., Chang, P., & Tanyolac, M. B. (2019). Genome-wide association studies of protein, lutein, vitamin C, and fructose concentration in wild and cultivated chickpea seeds. *Crop Science*, *59*, 2652–2666. https://doi.org/10.2135/cropsci2018.12.0738

Katuuramu, D. N., Hart, J. P., Porch, T. G., Grusak, M. A., Glahn, R. P., & Cichy, K. A. (2018). Genome-wide association analysis of nutritional composition-related traits and iron bioavailability in cooked dry beans (Phaseolus vulgaris L.). *Molecular Breeding*, *38*, 44. https://doi.org/10.1007/s11032-018-0798-x

Khazaei, H., Caron, C. T., Fedoruk, M., Diapari, M., Vandenberg, A., Coyne, C. J., McGee, R., & Bett, K. E. (2016). Genetic diversity of cultivated lentil (Lens culinaris Medik.) and its relation to the world's agro-ecological zones. *Frontiers in Plant Science*, *7*, 1–7, 1093. https://doi.org/10.3389/fpls.2016.01093

Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., Keurentjes, J. J., & van Eeuwijk, F. (2014). Marker-based estimation of heritability in immortal populations. *Genetics*, *199*, 379–398. https://doi.org/10.1534/genetics.114.167916

Kumar, J., Gupta, D. S., Kumar, S., Gupta, S., & Singh, N. P. (2016). Current knowledge on genetic biofortification in lentil. *Journal of Agricultural and Food Chemistry*, *64*, 6383–6396. https://doi.org/10.1021/acs.jafc.6b02171

Kumar, J., Sen Gupta, D., Baum, M., Varshney, R. K., & Kumar, S. (2021). Genomics-assisted lentil breeding: Current status and future strategies. *Legume Science*, *3*, 1–20. https://doi.org/10.1002/leg3.71

Kumar, H., Singh, A., Dikshit, H. K., Mishra, G. P., Aski, M., Meena, M. C., & Kumar, S. (2019). Genetic dissection of grain iron and zinc concentrations in lentil (Lens culinaris Medik.). *Journal of Genetics*, *98*, 1–14, 66. https://doi.org/10.1007/s12041-019-1112-3

Liber, M., Duarte, I., Maia, A. T., & Oliveira, H. R. (2021). The history of lentil (Lens culinaris subsp. culinaris) domestication and spread as revealed by genotyping-by-sequencing of wild and landrace accessions. *Frontiers in Plant Science*, *12*, 1–18, 628439. https://doi.org/10.3389/fpls.2021.628439

Long, W. (2015). *Automated amino acid analysis using an Agilent Poroshell HPH-C18 column*. Agilent Technologies, Inc.

Ma, Y., Marzougui, A., Coyne, C. J., Sankaran, S., Main, D., Porter, L. D., Mugabe, D., Smitchger, J. A., Zhang, C., Amin, M. N., Rasheed, N., Ficklin, S. P., & McGee, R. J. (2020). Dissecting the genetic architecture of Aphanomyces root rot resistance in lentil by QTL mapping and genome-wide association study. *International Journal of Molecular Sciences*, *21*(6), 2129. https://doi.org/10.3390/ijms21062129

Madurapperumage, A., Johnson, N., Thavarajah, P., Tang, L., & Thavarajah, D. (2022). Fourier-transform infrared spectroscopy (FTIR) as a high-throughput phenotyping tool for quantifying protein quality in pulse crops. *The Plant Phenome Journal*, *5*(1), e20047. https://doi.org/10.1002/ppj2.20047

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., & Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb)*, *108*, 285–291. https://doi.org/10.1038/hdy.2011.73

Manneberg, M., Lahm, H. W., & Fountoulakis, M. (1995). Quantification of cysteine residues following oxidation to cysteic acid in the presence of sodium azide. *Analytical Biochemistry*, *231*, 349–353. https://doi.org/10.1006/abio.1995.9988

Martín-Cabrejas, M. A., Aguilera, Y., Pedrosa, M. M., Cuadrado, C., Hernández, T., Díaz, S., & Esteban, R. M. (2009). The impact of dehydration process on antinutrients and protein digestibility of some legume flours. *Food Chemistry*, *114*, 1063–1068. https://doi.org/10.1016/j.foodchem.2008.10.070

Megazyme. (2019). Retrieved from November 29, 2023, from https://www.megazyme.com/protein-digestibility-assay-kit

Monsoor, M. A., & Yusuf, H. K. M. (2002). In vitro protein digestibility of lathyrus pea (Lathyrus sativus), lentil (Lens culinaris), and chickpea (Cicer arietinum). *International Journal of Food Science and Technology*, *37*, 97–99. https://doi.org/10.1046/j.1365-2621.2002.00539.x

Muthusamy, V., Hossain, F., Thirunavukkarasu, N., Choudhary, M., Saha, S., Bhat, J. S., Prasanna, B. M., & Gupta, H. S. (2014). Development of β-carotene rich maize hybrids through marker-assisted introgression of β-carotene hydroxylase allele. *PLoS ONE*, *9*, 1–22, e113583. https://doi.org/10.1371/journal.pone.0113583

Nair, B. M., Oste, R., Asp, N. G., & Dahlqvist, A. (1976). Enzymic hydrolysis of food protein for amino acid analysis. I. Solubilization of the protein. *Journal of Agricultural and Food Chemistry*, *24*, 386–389. https://doi.org/10.1021/jf60204a043

National Research Council Subcommittee on the Tenth Edition of the Recommended Dietary Allowances. (1989). Recommended dietary allowances. In *Recommended dietary allowances* (tenth ed.). National Academies Press. https://doi.org/10.1016/s0002-8223(21)22412-7

Papandreou, C., Becerra-Tomás, N., Bulló, M., Martínez-González, M. Á., Corella, D., Estruch, R., Ros, E., Arós, F., Schroder, H., Fitó, M., Serra-Majem, L., Lapetra, J., Fiol, M., Ruiz-Canela, M., Sorli, J. V., & Salas-Salvadó, J. (2019). Legume consumption and risk of all-cause, cardiovascular, and cancer mortality in the PREDIMED study. *Clinical Nutrition*, *38*, 348–356. https://doi.org/10.1016/j.clnu.2017.12.019

Patil, B., Lad, D., Bhagat, A., & Nawale, S. (2020). Assessment of biochemical parameters and genetic variability in chickpea (Cicerar arietinum L.) genotypes. *International Journal of Chemical Studies*, *8*, 1305–1308. https://doi.org/10.22271/chemi.2020.v8.i3r.9378

Pavan, S., Bardaro, N., Fanelli, V., Marcotrigiano, A. R., Mangini, G., Taranto, F., Catalano, D., Montemurro, C., de Giovanni, C., Lotti, C., & Ricciardi, L. (2019). Genotyping by sequencing of cultivated lentil (Lens culinaris Medik.) highlights population structure in the Mediterranean Gene Pool associated with geographic patterns and phenotypic variables. *Frontiers in Genetics*, *10*, 1–9, 872. https://doi.org/10.3389/fgene.2019.00872

Pfeiffer, B. (2017). *The improvement of grain sorghum productivity, black pericarp color, and protein digestibility*. https://hdl.handle.net/1969.1/173127

Pratap, A., & Kumar, J. (2011). *Biology and breeding of food legumes* (pp. 63–80). CABI. https://books.google.com/books?id=3Z-BW3V9nL8C. https://doi.org/10.1079/9781845937669.0000

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Rajendran, K., Coyne, C. J., Zheng, P., Saha, G., Main, D., Amin, N., Ma, Y., Kisha, T., Bett, K. E., McGee, R. J., & Kumar, S. (2021). Genetic diversity and GWAS of agronomic traits using an ICARDA lentil (Lens culinaris Medik.) reference plus collection. *Plant Genetic Resources: Characterisation and Utilisation*, *19*, 279–288. https://doi.org/10.1017/S147926212100006X

Ramsay, L., Koh, C. S., Kagale, S., Gao, D., Kaur, S., Haile, T., Gela, T. S., Chen, L. A., Cao, Z., Konkin, D. J., & Toegelová, H. (2021). Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. *bioRxiv*. https://doi.org/10.1101/2021.07.23.453237

Salaria, S., Boatwright, J. L., Thavarajah, P., Kumar, S., & Thavarajah, D. (2022). Protein biofortification in lentils (Lens culinaris Medik.) toward human health. *Frontiers in Plant Science*, *13*, 869713. https://doi.org/10.3389/fpls.2022.869713

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, *44*, 825–830. https://doi.org/10.1038/ng.2314

Semba, R. D., Ramsing, R., Rahman, N., Kraemer, K., & Bloem, M. W. (2021). Legumes as a sustainable source of protein in human diets. *Global Food Security*, *28*, 100520. https://doi.org/10.1016/j.gfs.2021.100520

Shekib, L. A. H., Zoueil, M. E., Youssef, M. M., & Mohamed, M. S. (1986). Amino acid composition and in vitro digestibility of lentil and rice proteins and their mixture (Koshary). *Food Chemistry*, *20*, 61–67. https://doi.org/10.1016/0308-8146(86)90167-6

Singh, U., & Jambunathan, R. (1981). Studies on desi and Kabull chickpea (Cicer arietinum L.) cultivars: Levels of protease inhibitors, levels of polyphenolic compounds and in vitro protein digestibility. *Journal of Food Science*, *46*, 1364–1367. https://doi.org/10.1111/j.1365-2621.1981.tb04176.x

South, A. (2011). Rworldmap: A new R package for mapping global data. *R Journal*, *3*, 35–43. Available at: http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf. https://doi.org/10.32614/RJ-2011-006

Tabangin, M. E., Woo, J. G., & Martin, L. J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, *3*, 5–8, S41. https://doi.org/10.1186/1753-6561-3-s7-s41

Talukdar, D. (2016). Exogenous thiourea modulates antioxidant defence and glyoxalase systems in lentil genotypes under arsenic stress. *Journal of Plant Stress Physiology*, *2*, 9. https://doi.org/10.19071/jpsp.2016.v2.3041

Thavarajah, D., Lawrence, T., Boatwright, L., Windsor, N., Johnson, N., Kay, J., Shipe, E., Kumar, S., & Thavarajah, P. (2023). Organic dry pea (Pisum sativum L.): A sustainable alternative pulse-based protein for human health. *PLoS ONE*, *12;18*(4), e0284380. https://doi.org/10.1371/journal.pone.0284380 PMID: 37043476; PMCID: PMC10096185

The Good Food Institute. (2023). Plant-based market overview. Retrieved August 17, 2023, from https://www.gfi.org/marketresearch

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome*, *14*, 1–17, e20077. https://doi.org/10.1002/tpg2.20077

U.S. Department of Agriculture. (2019). Lentils, raw. Food data central, NDB # 16069. Retrieved March 2023, from https://fdc.nal.usda.gov/fdc-app.html#/food-details/172420/nutrients

Wang, N., & Daun, J. K. (2006). Effects of variety and crude protein content on nutrients and anti-nutrients in lentils (Lens culinaris). *Food Chemistry*, *95*, 493–502. https://doi.org/10.1016/j.foodchem.2005.02.001

Wang, J., & Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics*, *19*, 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, Y., Zhang, L., & Zhu, S. (2014). 1-Methylcyclopropene (1-MCP)-induced protein expression associated with changes in Tsai Tai (Brassica chinensis) leaves during low temperature storage. *Postharvest*

Biology and Technology, *87*, 120–125. https://doi.org/10.1016/j.postharvbio.2013.08.016

Wickham, H. (2011). The Split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*, 1–29. Available at: https://www.jstatsoft.org/v40/i01/. https://doi.org/10.18637/jss.v040.i01

Wickham, H., & Wickham, H. (2016). *Data analysis*. (pp. 189–201). Springer International Publishing.https://ggplot2.tidyverse.org

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Johnson, N., Boatwright, J. L., 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*(3), 640–655. https://doi.org/10.1002/ppp3.10470