


Genotypic Contrasting of Protein and Flavonoid Contributes to Differential Responses of Targeted Metabolites in Soybean Seeds



Dinh Ha Tran³, Trong Dai Tran¹, Tien Dung Nguyen⁴, Xuan Vu Nguyen⁴, Xuan Binh Ngo⁴, Duong Van Doan², Anh Tuan Tran⁵, Thi Phuong Anh Dang⁶, Thi Thao La¹⁰, Van Tinh Nguyen⁷, Minh Nguyen^{8,f}, Van Hung Hoang⁹ and Van Hien La^{1,2,*} 

¹Institute of Life Science, Thai Nguyen University, Quyet Thang, Thai Nguyen 250000, Vietnam

²Center of Crop Research for Adaptation to Climate Change, Thai Nguyen University of Agriculture and Forestry, Quyet Thang, Thai Nguyen 250000, Vietnam

³Faculty of Agronomy, Thai Nguyen University of Agriculture and Forestry, Quyet Thang, Thai Nguyen 250000, Vietnam

⁴Institute of Biotechnology and Food Technology, Thai Nguyen University of Agriculture and Forestry, Quyet Thang, Thai Nguyen 250000, Vietnam

⁵Vietnam National University of Agriculture, Gia Lam, Hanoi 12406, Vietnam

⁶Department of Agronomy Centre for Flower, Ornamental Research and Development, Fruit and Vegetable Research Institute, Vietnam National University of Agriculture, Vietnam Academy of Agriculture Science, Gia Lam, Hanoi 131000, Vietnam

⁷Applied Biomedical Research Institute, Buon Ma Thuot Medical University, Buon Ma Thuot 630000, Vietnam

⁸Institute of Earth Science, Academia Sinica, Taipei City, Taiwan

^fPresent Address: Hanoi School of Business and Management (HSB), Vietnam National University, Hanoi (VNU), B1, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

⁹Bonic Science, Thai Nguyen University, Tan Thinh, Thai Nguyen 250000, Vietnam

¹⁰Department of Biotechnology, Vietnam National University of Agriculture, Hanoi, Vietnam

Abstract:

Introduction/Objective: Soybean is a major source of various nutrients. Increasing demand for soybeans has created considerable impetus for exploring the nutritional quality of soybeans. We aimed to collect soybean varieties rich in nutrients.

Materials and Methods: Metabolite analysis was carried out for seed compositions, including protein, phenolics, and flavonoids, along with gene expression of protein and phenolic metabolism-related enzymes in 10 soybean accessions collected from different geographical regions.

Results: Total protein content ranged from 29.7% to 35.7%, depending on soybean germplasm accessions. Among them, Vang Ha Giang (VHG) exhibited relatively high protein content, while Cuc Vo Nhai (CVN) had comparatively low protein content. Further analysis of seed compounds indicated that the phenolic compounds were higher in cultivars Dau Tuong Den (DTD) and CVN, with a total phenolic content of 37.7 $\mu\text{g g}^{-1}$ and total flavonoid of 2.1 mg g^{-1} . These results were reinforced by analysis of gene expression levels of candidate genes β -conglycinin (7S) and glycinin (11S) involving protein storage, phenylalanine ammonia-lyase 1 (*GmPAL1*) and chalcone synthase 8 (*GmCHS8*) genes related to phenolic and flavonoid synthesis, which showed similar correlation. We revealed that protein content was correlated with seed weight but not with seed color, even though significant variations were found among soybean genotypes, while flavonoid was affected by seed coat color. Furthermore, the negative correlation of protein with flavonoids demonstrated intricate relationships among seed components.

Conclusion: Protein and flavonoid alteration in seeds is subject to major-effect-genotypes in landrace and breeding cultivar selection, and genotype variants are relevant to geographical regions. Our study provides intricate relationships among seed nutritional components and offers insight into the alteration of soybean quality.

Keywords: Black soybean, Flavonoid, Legume, Protein, Soybean, Seed quality.

© 2025 The Author(s). Published by Bentham Open.

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: <https://creativecommons.org/licenses/by/4.0/legalcode>. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



Received: November 04, 2024

Revised: November 24, 2024

Accepted: November 26, 2024

Published: February 14, 2025



Send Orders for Reprints to
reprints@benthamscience.net

*Address correspondence to this author at the Institute of Life Science, Thai Nguyen University, Quyet Thang, Thai Nguyen 250000, Vietnam and Center of Crop Research for Adaptation to Climate Change, Thai Nguyen University of Agriculture and Forestry, Quyet Thang, Thai Nguyen 250000, Vietnam; E-mail: hiencnsh87@gmail.com

Cite as: Tran D, Tran T, Nguyen T, Nguyen X, Ngo X, Van Doan D, Tran A, Dang T, La T, Nguyen V, Nguyen M, Hoang V, La V. Genotypic Contrasting of Protein and Flavonoid Contributes to Differential Responses of Targeted Metabolites in Soybean Seeds. *Open Agric J*, 2025; 19: e18743315365636. <http://dx.doi.org/10.2174/0118743315365636241226060752>

1. INTRODUCTION

Soybean (*Glycine max* L.) is a primary crop in agriculture and provides a major nutrient for humans [1]. It has been well known as a crop with high amounts of proteins, amino acids, and oil, as well as various bioactive compounds [1-3]. The different bioactive compounds in soybean seeds are associated with human health, including antioxidant activities, cholesterol-lowering effects, and anti-obesity activities [1, 4]. Black soybean contains phenolics, isoflavones, and anthocyanins. These compositions have been known to contribute to various health benefits [4-6]. Therefore, the composition of seeds has a significant impact on the quality of soybean products.

Soybeans have been widely used in food because of their characteristic richness in nutrients (protein, oil, fatty acid, soluble sugar, and isoflavone). The variation of nutrition is primarily influenced by factors, such as the genetic background of a cultivar, location, climate, and major group [6-9]. These factors affect the arrangement and shifts of targeted nutrients within soybean seeds by modifying the metabolic phenylpropanoid pathway [6, 10], flavonoid synthesis [6, 11], and amino acid metabolism [12]. Evidence showed that the differences in the genetics of cultivars contributed to 83% of the variation in protein contents [12]. With the diversity of species composition, there are differences in morphological characteristics response, nutritional composition, and quality. Recent research showed that the content of antioxidants depends on the characteristics of soybean varieties and ecological growing regions [13-15]. Soybean cultivars have been characterized as two distinct species, *Glycine max* and *Glycine soja*, based on their genetics and habitats [16]. *Glycine max* has been known as an essential cultivated soybean in the world, providing higher quality proteins and isoflavones compared to *Glycine soja* (a wild species), and it has been widely used in the food industry [2].

It is noteworthy that the screening of a substantial collection of germplasms is essential to aid the identification of genotypes with elevated levels of protein and bioactive compounds for genetic improvement programs. Nevertheless, comprehensive information regarding the profiles and content of protein and bioactive compounds within the seeds of soybeans from a diverse set of Vietnamese soybean germplasms originating from various ecoregions remains limited. Recently, we have

identified merely two studies that outlined and assessed variations of protein content in 321 soybean (*Glycine max* L.) varieties [12], along with 218 Chinese soybean strains and 115 wild soybean samples from the USA [7]. As previously mentioned, other research has similarly emphasized the influence of genotype and environmental conditions on flavonoid content in wild black soybeans and cultivated black soybean genotypes [6].

Several studies have documented the fluctuation of biochemical or metabolic constituents, and the connection between them contributed to the yield and quality of soybean seeds. Various qualitative traits of soybean seeds indicate that seed coat color could be one of the key factors in soybean seed quality [17]. The level of primary and secondary metabolites is influenced by seed size, where the larger seeds contain higher amounts of starch, sugar, protein, and lipids compared to smaller seeds [18-20]. For example, the size of soybean seed showed a favorable association with oleic acid levels, while it exhibited a negative correlation with linoleic acid levels [18, 21]. On the other hand, seed coat coloration influenced isoflavones, fatty acids, and flavonoid content in soybean [6, 22, 23]. However, seed coat pigmentation did not influence the inverse correlation of protein and lipid concentrations [24]. It has been reported that the variation of level and content of metabolites are highly interdependent in soybean seeds. Among these, there is a complex inverse relationship between protein, oil, and sugar in soybean seeds. Zhang *et al.* (2018) [12] found a strong inverse relationship between oil and protein, indicating that as one increases, the other decreases, which contrasts with sucrose levels targeted for breeding high-yield soybean seeds. Lee and Son (1993) [25] found a positive correlation between oil and protein content among 1,087 colored soybean accessions. Interestingly, an inverse correlation of protein and oil content was observed in both whole soybean seeds and seed coats [26]. Additionally, the levels of protein were negatively correlated with stachyose or raffinose in 43 soybean progenies [27].

Nevertheless, little is known about how target metabolites, including protein, phenolic, and flavonoid, respond to seed coat colors and other physiological properties, such as relative seed diameter, across a diversity of soybean seed germplasm accessions. In this study, various soybean germplasm entries with a range of seed coat colors were collected locally to assess the

relationship between seed coat color and seed size on the inter-relationships and variations of prioritized metabolites. Hence, the current study aimed to evaluate the conjecture that focused metabolites (protein, phenolic, and flavonoid) would be impacted in various ways by seed coat colors and seed dry weight and explore how the variation of these compounds could influence the relationship or correlation between protein, phenolic, and flavonoid in diverse soybean seed germplasm accessions. Additionally, geographical maps were utilized to identify hotspot areas, where cultivars showed elevated levels of seed nutritional components.

2. MATERIALS AND METHODS

2.1. Plant Material and Field Trials

A population of 10 soybean germplasm accession of plant introductions (PIs), including Vietnamese landraces and cultivars of soybean species, was used in this study (Table S1). The PIs were randomly collected from the Plant Resource Center (<http://prc.org.vn>) and the mountainous provinces in northern Vietnam.

The field experiments were performed at the experimental grounds of the Thai Nguyen University of Agriculture and Forestry (21°33'51" N, 105°52'46" E) located in Thai Nguyen City, Vietnam. The sandy, loamy soil was detected by pH 6.5 – 7, 0.6% OM, 0.06% total N, and 2.5 dSm⁻¹ EC. Field trials were conducted from February to June, 2022. The ridges were created with diameters of 30 cm height and 60 cm width at the base using an FJ601 tiller (Fuji, Japan) coupled to an RA3 rotary. The thickness of the topsoil horizon was characterized by organic matter accumulation exceeding 30 cm in the plots. The experimental design included two areas featuring flat and ridged plots arranged in a completely randomized block format. In each area, three plots were designated for repetition analysis. Each soybean seed was planted at 20 cm intervals, with 30 plants cultivated per plot. For three days following sowing, water was sprayed daily to encourage seed germination. The commercial nitrogen-phosphate-kali (NPK, 10N: 5 P₂O₅: 5 K₂O) fertilizers were used in the test field two times during the experiments at the 3-leaf stage and 7-leaf stage of the plant development process.

2.2. Phenotypic Evaluation

Phenotypic data were indicated by collecting eight morphological and physiological traits, including Times Cultivation (TC) plant height at physiological maturity (PH), Leaf Area Index (LAI), number of pods per plant (NPP), Seed Yield (SY), and 100-Seed Weight (SW). In each plot, the morphological and physiological traits of five randomly taken plants were recorded for observation. The data for these was collected on a plot basis.

The flower and seed phenotypic experiments were performed using a Light-Stemi 508 microscope (Carl Zeiss, Germany). Images of the flower and seed were captured using a CCD camera at 4 × magnification. The number and length of the lateral roots were measured 1 cm from the primary root using ImageJ software [28].

2.3. Total Protein in Seeds Analysis

Protein content in seeds was estimated by Bradford reagent kit B6916 (<https://www.sigmaldrich.com/>) based on bovine serum albumin (BSA) as standard protein. The protein profile was extracted by 50 mM potassium phosphate buffer (pH = 7.5). The absorbance of the mixture was measured using the Synergy H1 Hybrid Reader (Biotek) at 725 nm. The total protein content was expressed in terms of the standard curve of BSA equivalents per gram of fresh-weight extract. The protein composition was normalized, allowing for the quantification of protein proportion.

2.4. Total Phenolics, Hydroxycinnamic Acid, and Flavonoids Analysis

Total phenolic content was assessed by using the Folin-Ciocalteu method [29]. In brief, 200 mg of fresh leaves were extracted by using 80% methanol. A colorimetric reaction of the mixture was conducted using 2N Folin-Ciocalteu reagent, followed by the addition of 20% Na₂CO₃ and allowed to react for 60 minutes. Absorbance at 725 nm was measured with a Synergy H1 Hybrid Reader (Biotek, South Korea). The total phenolics were measured using the calibration curve based on gallic acid equivalents per gram of fresh weight extract.

Total hydroxycinnamic acids (THA) content in seeds was determined according to the method outlined by Štefan *et al.* (2014) [30]. After reacting with Aron's reagent (NaNO₂-Na₂MoO₄, 1:1), HCl, and NaOH, the absorbance at 505 nm was documented and determined using a chlorogenic acid standard curve.

The total flavonoid content in seeds was determined by the aluminum chloride colorimetric method [31], with minor modifications. In brief, 200 mg of dry seeds were extracted using 80% methanol. The crude extract was diluted with 0.5 mL of distilled water and incubated with the addition of 75 µL of 5% NaNO₂ and 0.3 mL of 10% AlCl₃. After incubation for 5 min, 1 M NaOH (0.5 mL) was added to the mixture, and absorbance was recorded at 510 nm. The total phenolic content was estimated by using a calibration curve based on quercetin equivalents per gram of dry seed weight extract.

2.5. Phenolics and Flavonoids Quantification by RP-HPLC

The content of individual phenolics and flavonoids in the seeds was quantified following the procedure detailed by Das *et al.* (2018) [32]. Briefly, 100 mg of soybean powder was extracted in 80% methanol. The collected supernatant was filtered using a 0.2 µm PVDF syringe filter. Soluble phenolics and flavonoids were analyzed by a system coupled with a UV-VIS detector (SPD-20A, Shimadzu, Kyoto, Japan). Chromatographic separation was achieved by a Spherisorb® ODS2 column with dimensions

of 4.60 × 250 mm (Waters, Milford, MA, USA) at 30°C. Orthophosphoric acid (0.1%) in water (v/v) (eluent A) and methanol (v/v) (eluent B) were employed as the mobile phases. The retention time and regression equation of the standards were used for the determination of the soluble phenolic and flavonoid content.

2.6. RNA Extraction and Quantitative Real-time PCR Analysis

Total RNA was isolated from 100 mg of seeds using a cell lysis buffer, following a previously established DNA-free RNA isolation method [33]. The DNA-free RNA was used as the template for cDNA synthesis by using the GoScript Reverse Transcription System in accordance with the manufacturer's instructions (Takara, DALIAN, Japan). Quantitative reverse transcription-PCR (qRT-PCR) was conducted on a LightCycler 96 real-time PCR system using 2X SYBR Green qPCR Master Mix (Takara, DALIAN, Japan). RT-PCR reaction 25 µL included 12.5 µL 2X SYBR Green Master Mix, 1 µL forward primer (10 pmol), 1 µL reverse primer (10 pmol), and 1 µL cDNA, and 9.5 µL water-free DNase. RT-PCR cycles contained an early denaturation at 95 °C for 5 minutes followed by 35 cycles of denaturation at 95 °C for 30 seconds, annealing at T_m for 30 seconds, extension at 72 °C for 30 seconds, and a final expansion at 72 °C for 5 min. T_m of the gene-specific primers utilized for qRT-PCR was conducted in duplicate for each of the three separate samples (Supplementary Table S4). Relative gene expression levels were determined and inferred from the threshold value (C_t), and the actin gene was utilized as an internal control. For the quantification of the relative transcript levels, we used the $2^{-\Delta\Delta C_t}$ method [34].

2.7. Metabolic Analysis

To further investigate the functional interpretations and associations of the identified metabolites compounds in soybean seeds, Principal Component Analysis (PCA), one-way ANOVA visualization, Radom Forest classification, and interrelation interactions among biochemical substances like protein, phenolics, and flavonoids were generated using MetaboAnalyst 5.0 (<http://www.metaboanalyst.ca>).

2.8. Geographical Distribution Mapping

Geographical distribution maps of soybean seed nutrition were created using QGIS 10.0 (<https://www.qgis.org/en/site/>) using ordinary interpolation [35]. QGIS is commonly used in geographic information system (GIS) applications because it allows easy map creation, geographic data compilation, and spatial data management. To create the maps in this study, we applied interpolation to the mean seed nutrition data across various locations and cultivars, incorporating geographical factors (longitude and latitude). This method allocates weights to known sample points to predict the values of unknown sample points.

2.9. Statistical Analysis

The current study employed a fully randomized design involving three repetitions for every treatment and collection date. Analysis of variance (ANOVA) was conducted on whole data, and Duncan's multiple range test was applied to compare the means of each replicate for every sampling time. All statistical analyses were conducted using SAS 9.1 (SAS Institute, Inc., 2002-2003), with differences regarded as significant at $p < 0.05$.

3. RESULTS

3.1. The Response of Protein to Seed Weight and Seed Coat Color in Various Soybean Cultivars

The morphology of soybean varieties is shown in Fig. (1), and the phenotype of plants at the maturity and harvest stages clearly showed differences in plant height, architecture of plant, degree of branching, and flower and seed colors (Table 1). On the other hand, the protein content of soybean seeds is crucial for assessing their nutritional value and quality. In the soybean cultivar, the protein content varied from 29.7% to 35.2% (Fig 2). In this study, protein, one of the targeted metabolites, was determined to be affected by seed weight. This response was confirmed by the correlation scoring plot (Fig. S1), which showed that protein content was higher in larger seeds compared to smaller ones, a finding that was also reported in a previous study [23]. Seed size (measured as the dry weight of 100 seeds) significantly affected the protein content, making it an important plant growth characteristic.

Table 1. Plant physiological characteristics in soybean landraces and cultivars.

Cultivars	Time Cultivation (day)	Plant Height (cm)	Leaf Area (cm ²)	Fruit per Plant	Seed Weight (g)	Color
DT84	95.0±0.94 ^b	51.0 ± 1.66 ^{cd}	7.2 x 4.5	83.5 ± 1.66 ^b	0.223 ± 0.011 ^a	Yellow
DTD	92.0±0.47 ^{bc}	74.0 ± 0.95 ^b	6.8 x 4.3	93.0 ± 1.61 ^b	0.100 ± 0.007 ^{cd}	Black
VMK	93.5±0.23 ^b	48.5 ± 1.06 ^d	6.8 x 4.8	69.0 ± 7.59 ^c	0.209 ± 0.008 ^{ab}	Yellowish
VCB	103.0±0.47 ^{ab}	89.0 ± 6.01 ^a	6.4 x 4.3	42.0 ± 6.17 ^d	0.178 ± 0.002 ^b	Yellowish
VHG	105.0±0.48 ^a	84.0 ± 1.90 ^{ab}	6.2 x 5.1	65.0 ± 5.22 ^c	0.118 ± 0.004 ^c	Yellow
VQN	94.5±0.24 ^b	73.5 ± 4.98 ^b	7.3 x 6.6	96.0 ± 0.24 ^b	0.109 ± 0.006 ^c	Yellow
CHBD1	101.0±1.02 ^{ab}	81.3 ± 3.54 ^{ab}	9.3 x 7.6	152.0 ± 7.12 ^a	0.087 ± 0.005 ^d	Yellowish
CHLS	92.0±0.47 ^{bc}	66.0 ± 4.24 ^c	4.8 x 3.8	88.0 ± 1.90 ^b	0.118 ± 0.008 ^c	Yellow

(Table 1) contd.....

Cultivars	Time Cultivation (day)	Plant Height (cm)	Leaf Area (cm ²)	Fruit per Plant	Seed Weight (g)	Color
CVN	91.0±0.48 ^c	77.5 ± 4.98 ^b	4.8 x 3.3	39.0 ± 2.37 ^d	0.097 ± 0.002 ^{cd}	Yellow
DTSM	92.5±0.71 ^{bc}	66.5 ± 4.51 ^c	7.2 x 5.9	68.0 ± 2.37 ^c	0.101 ± 0.004 ^{cd}	Yellow

Note: Different lowercase letters in a column indicate significant differences at $P < 0.05$ according to the Duncan's multiple range test.

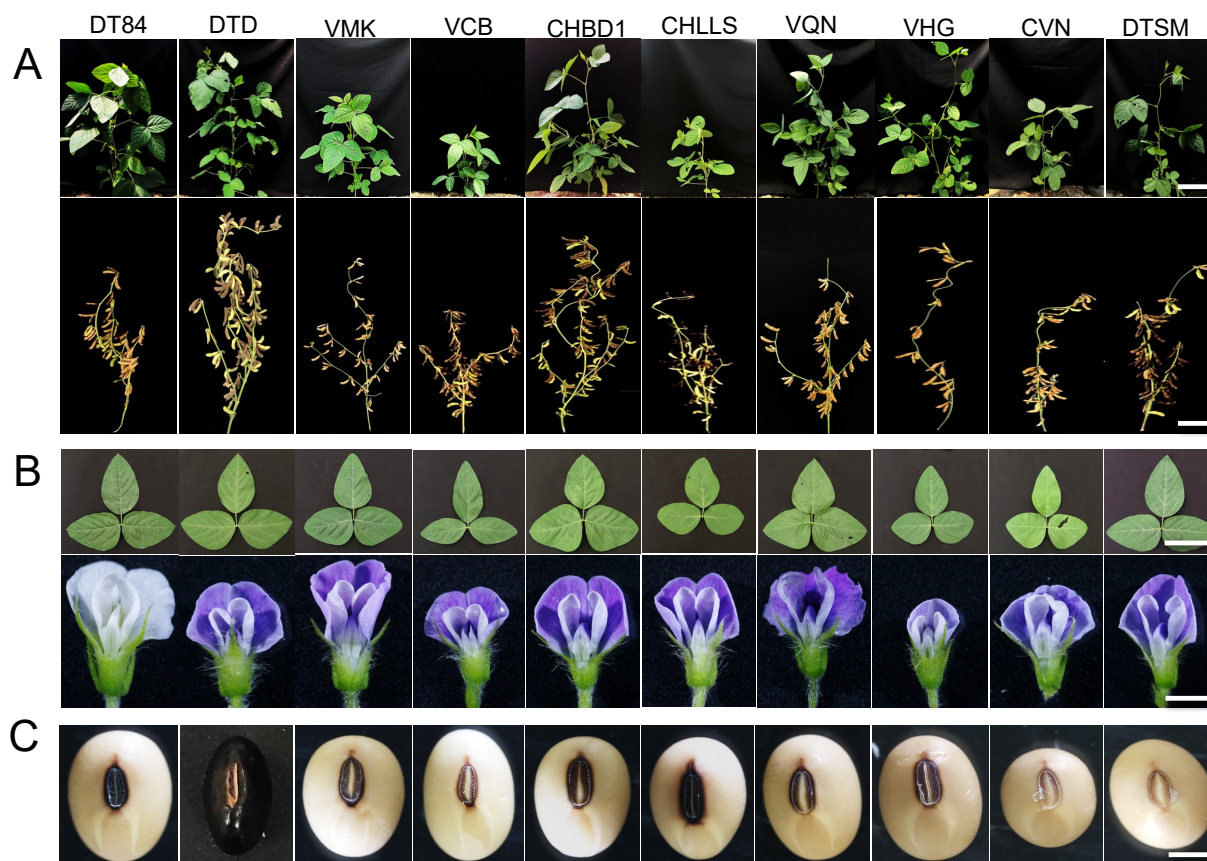


Fig. (1). Phenotypic variations of soybean cultivars. **(A)** Phenotypic of plants at the maturity and harvest stages. **(B)** Leaves and flower collected from maturity state. Leaves are selected at leaf number 5 (No.5) from root of the tree to the top. Flower was collected after 2 days opened. The soybean flower colors are most violet in all of landrace cultivars and DTD cultivar, whereas DT84 cultivar has white-colored flower. The photo was taken under a Stemi 508 microscope (Carl Zeiss, Germany) with a 0.5X, scan bar (2,0 cm). **(C)** Seeds size and color of the different cultivars. Seed sizes are small, medium, and large. All most cultivars represent light-yellow to dark-yellow seed, except for DTD black color-coded accession. The photograph was taken under a Stemi 508 microscope (Carl Zeiss, Germany) with a 1X scan bar of 1 cm.

Protein content was categorized into three groups based on total protein content: (i) Group 1, with four cultivars had a total protein $< 30\%$ including Cuc Vo Nhai (CVN); (ii) Group 2 had a total protein ranging from $\geq 30\%$ to $< 35\%$, including seven soybean cultivars DT84, Dau Tuong Den (DTD), Vang Muong Khuong (VMK), and Vang Cao Bang (VCB), Vang Quang Ninh (VQN); Cuc Huu Lung Lang Son (CHLLS), Dau tuong Song Ma (DTSM); and Group 3 included two cultivars that had total protein $\geq 35\%$, including Vang Ha Giang (VHG) and Cuc Ha Bac Dang 1 (CHBD1). Two cultivars, CHBD1 and VHG, exhibited the highest protein storage. The storage of protein in soybean seed (*Glycine max* (L.) Merr. cultivars) correlated to two major storage proteins, a glycosylated

7S protein (conglycinin) and a non-glycosylated 11S protein (glycinin) [36, 37]. Depending on the genotypes, these genes were highly expressed during the soybean seed maturity process. β -conglycinin (7S) and glycinin (11S) released on protein storage in soybean seeds (Meinke *et al.* 1981) exhibited 4.8-fold and 3.2-fold higher expression in VHG compared to the DT84 and other soybean cultivars (Fig. 1B and C). The soybean cultivars featured in this study presented a wide range of protein content, and these landraces are rich sources of protein. It is worth mentioning that these soybean cultivars revealed some landrace soybean cultivars that are mainly consumed as a food source with a potential of high protein content in seeds.

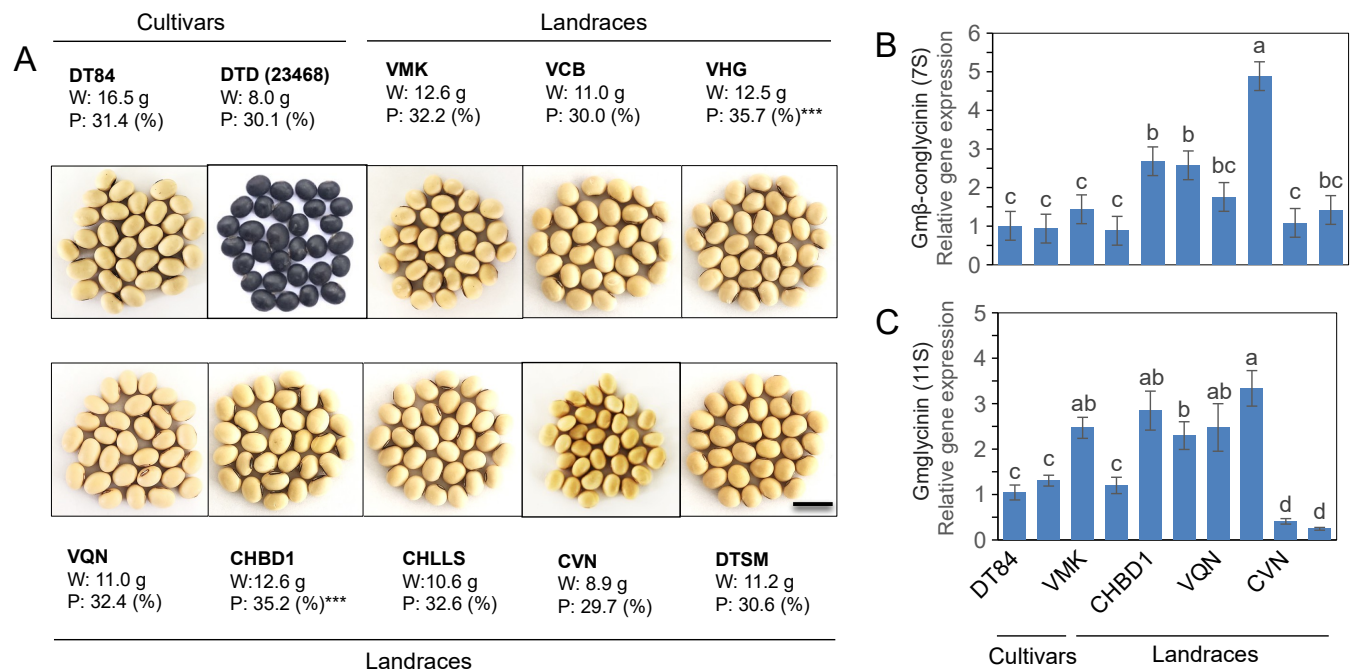


Fig. (2). Variation protein (P) contents and seed weight (W) in soybean seeds of landrace and cultivars accessions. Soybean seed weight was estimated by 100 seeds.

(A) The different colors of seeds derived from yellow in all of soybean landraces and cultivars. Relative gene expression regulation of protein, phenolic and flavonoid storage in soybean seeds.

(B) Gmconglycinin (Glycosylated 7S protein) and (C) Gmglycinin (non-glycosylated 11S protein) are the two major storage proteins in soybean seeds. Scan bars 1 cm. The data represent mean ± SE (n = 3). Asterisks indicate significant differences between the different soil moisture percentages as determined by Duncan’s t-test;

***p < 0.001.

3.2. Total Phenolic and Flavonoid Concentration Affected Seed Weight and Seed Colors

It has been found that the bioactive compounds of soybean seed mainly consist of total hydroxycinnamic acid (THA), phenolic acid (PA), and flavonoid (FA). Among these, THA has been known as a precursor in the biosynthesis process of phenolic and flavonoid (Fig. 3A). Consequently, large variations in the content of THA, PA, and FA, respectively, ranging from 2.0 - 13.0 (µg g⁻¹), 11.6 - 22.0 (µg g⁻¹), and 0.3 - 2.0 (µg g⁻¹), have been observed (Fig. 3B-D). The result targeted the conjunction of phenolic and flavonoid distribution, 100 seeds dry weight, and seed pigmentation, as displayed in Fig. (3). It was found that the majority of yellow soybean accessions have a higher dry mass of 100 seeds compared to black soybean seed accessions. However, the dry weight of 100 seeds from soybean germplasm accessions significantly influenced seed coat color. Two cultivars, DTD and CVN, had the lowest seed weight, but these cultivars showed higher total flavonoid content compared to other cultivars

(Fig. 3B-D). Intriguingly, seed coat colors tended to be highly responsive to targeted metabolites when comparing black soybean and yellow soybean seeds. However, the comparison of soybean seeds with similar seed coat colors indicated lower variation in the content of targeted metabolites, although the significance varied slightly among different shades of seed coat colors (Fig. 2 and 3B-D). Furthermore, seed coat color appeared to have a greater impact on the variation of bioactive compounds in yellow soybean seeds compared to black soybean seeds. An accumulation of metabolite compounds was found to be related to their maker, which is associated with most seed storage compounds. Two candidate genes, *phenylalanine ammonia-lyase 1* (GmPAL1) and *chalcone synthase* (GmCHS8), encoding putative for phenolic and flavonoid synthesis have also been identified in soybean seeds [10, 38]. Indeed, two genes, GmPAL1 and GmCHS8, were upregulated in cultivars DTD (2.3-fold and 14.2-fold) and CVN (1.5-fold and 5.1-fold), respectively (Fig. 3E-F), which are substantial for the phenolic and flavonoid accumulation in soybean seeds.

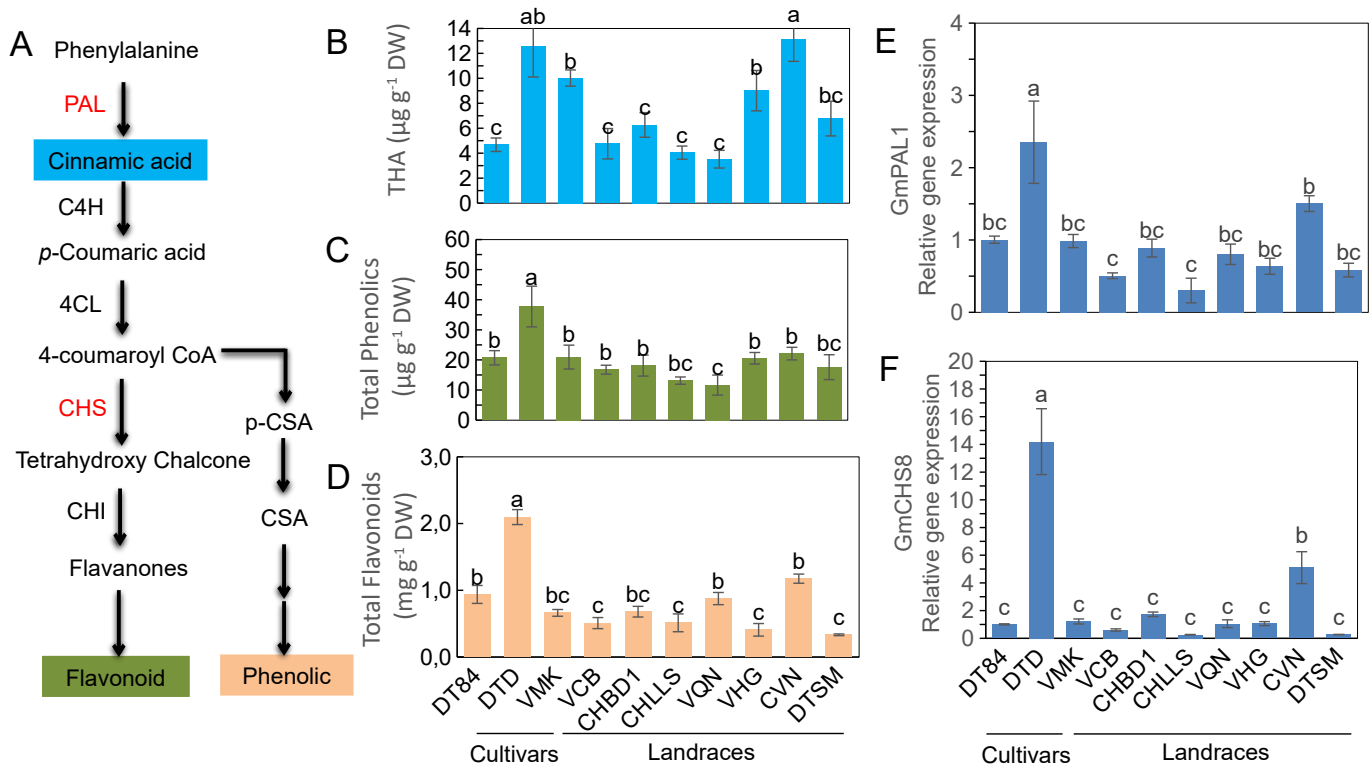


Fig. (3). Seed quality characterize in seeds of soybean landrace and cultivars accessions.

(A) The metabolic pathway synthesizes of phenolic and flavonoid derived from phenylpropanoid pathway. The targeted metabolites indicate by cinamate (hydroxycinnamic acid -THA, blue color), phenolic acid (PA, green color), and flavonoid (FA, light-orange color).

(B) Hydroxycinnamic acid -THA.

(C) Total phenolic acid.

(D) Total flavonoid. Relative gene expression regulation of phenolic and flavonoid storage in soybean seeds.

(E) Phenylalanine ammonia lyase 1 (*GmPAL1*) and (F) chalcone synthase (*GmCHS8*) encoding putative for phenolic and flavonoid synthesis identified in soybean seeds. The data represent mean \pm SE (n = 3).

Data with different letters in a vertical column are significantly different at $p < 0.05$ according to Duncan's multiple range test.

3.3. Cluster Targeted Metabolites Response to Seed Weight and Seed Coat Colors

To comprehend the contributions of specific and overall targeted metabolites (protein, THA, phenolic, and flavonoid) to designated seed weight and seed coat color, we conducted cluster analysis revealing the co-function of targeted metabolites, as shown in Fig. (4). The heatmap indicated that seed coat color could not affect the co-fluctuation of targeted metabolites, as the categorization of seed coat colors was largely uniformly distributed based on their arrangement. Nevertheless, black soybean seeds were closely clustered primarily around total phenolic content in two areas (seed weight and seed coat colors),

while the remaining black soybean seeds were distributed evenly across various cultivars. Meanwhile, the metabolic clustering was primarily divided into two distinct clusters for each factor analysis of seed weight or seed coat color. The first cluster consisted of the positive correlation of total protein to seed weight, while phenolic compounds were negatively correlated to seed weight. In the second cluster, total flavonoid, THA, and total phenolic had a positive correlation clustered in seed coat colors but *vice versa* in protein, respectively. Interestingly, total phenolic was closely linked in both clusters (Fig. 4A-B). Furthermore, the ANOVA results indicated that protein and flavonoid have an inverse correlation in seed weight or seed color (Fig. 4C-D).

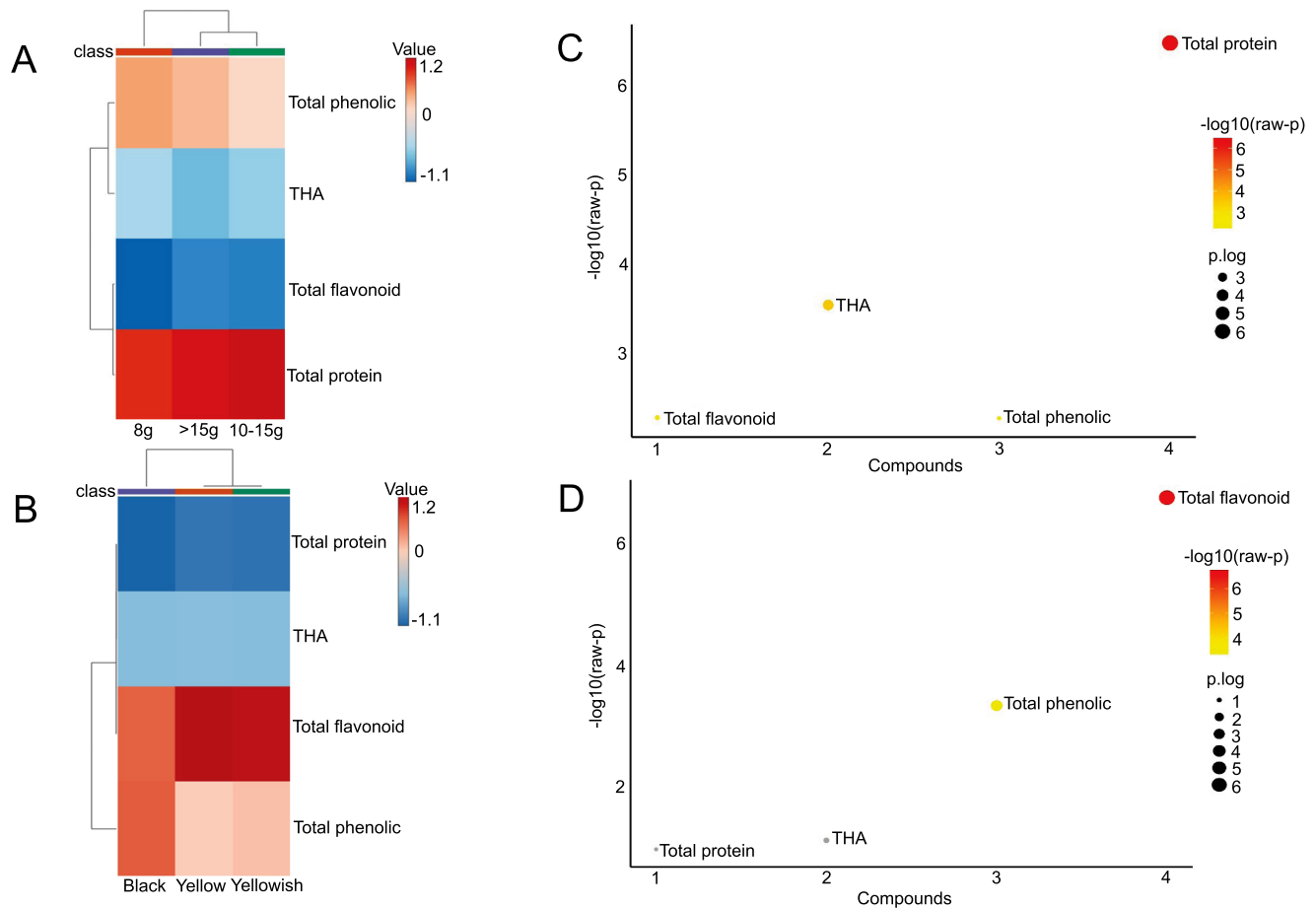


Fig. (4). Cluster the seed weight and seed coat colors effects on targeted metabolites in seeds of soybean accessions. **(A)** Seed weight effect on targeted metabolites. **(B)** Seed coat color effect on targeted metabolites. The targeted metabolites characterized in Table 2. **(C)** and **(D)** ANOVA analysis the significant or insignificant effect of seed weight and seed coat colors on total protein, THA, total phenolic acid, and total flavonoid. One-way ANOVA & post-hoc Tests. ANOVA p-value (FDR) cutoff: 0.05. Red and yellow spots have a positive correlation, while grey spots have a negative correlation.

3.4. Targeted Metabolite association to Seed Coat Colors

A more detailed analysis of the distribution of the targeted metabolites (protein, THA, phenolic, and flavonoid) in seed nutrition compounds was almost entirely based on the biochemical profile. Targeted metabolite responses from yellow to black soybean seeds were highly distributed by phenolic compounds. ANOVA analysis indicated substantial differences ($p < 0.001$) in order of total flavonoid, protein, THA, and PA contribution to yellow to black seed coat colors, respectively (Fig. 5). Among these, total flavonoid content was strongly associated with yellow to black seed colors, whereas total phenolic content showed a weaker association with seed coat color. Moreover, seed coat color was relatively unaffected by protein content, and total flavonoid content varied between yellow and black seed coat colors. As a result of one-way ANOVA analysis, FA, a secondary metabolite, demonstrated a significant difference, whereas

proteins, THA, and PA have no significant differences between accessions from disparate ecoregions and cultivars (Fig. 5A). This result indicated that the variation of FA and its related metabolites was abundant among seed coat colors in different cultivars. The coefficient score value (CV) indicated that the greatest contribution came from FA (> 0.07), followed by THA (> 0.05) and protein (0.02), whereas PA (< 0.01) provided the least contribution (Fig. 5B). The findings from PCA showed that the initial two components (PC1 and PC2) explained 83.6% of the total variation noted (Fig. 5C). The cultivars of soybean were grouped into two PCAs based on the content of the variation of seeds nutrition content. The Random Forest classification was performed to delve deeper into the metabolites that influenced the dissimilarity in soybean cultivars. The highest error ranking of cultivar was clustered with VMK, VQN, CHBD1, VHG, VCB, and DTSM, followed by DT84 and CHLLS, and the lowest clustered was with DTD and CVN (Fig. 5D and Table S2).

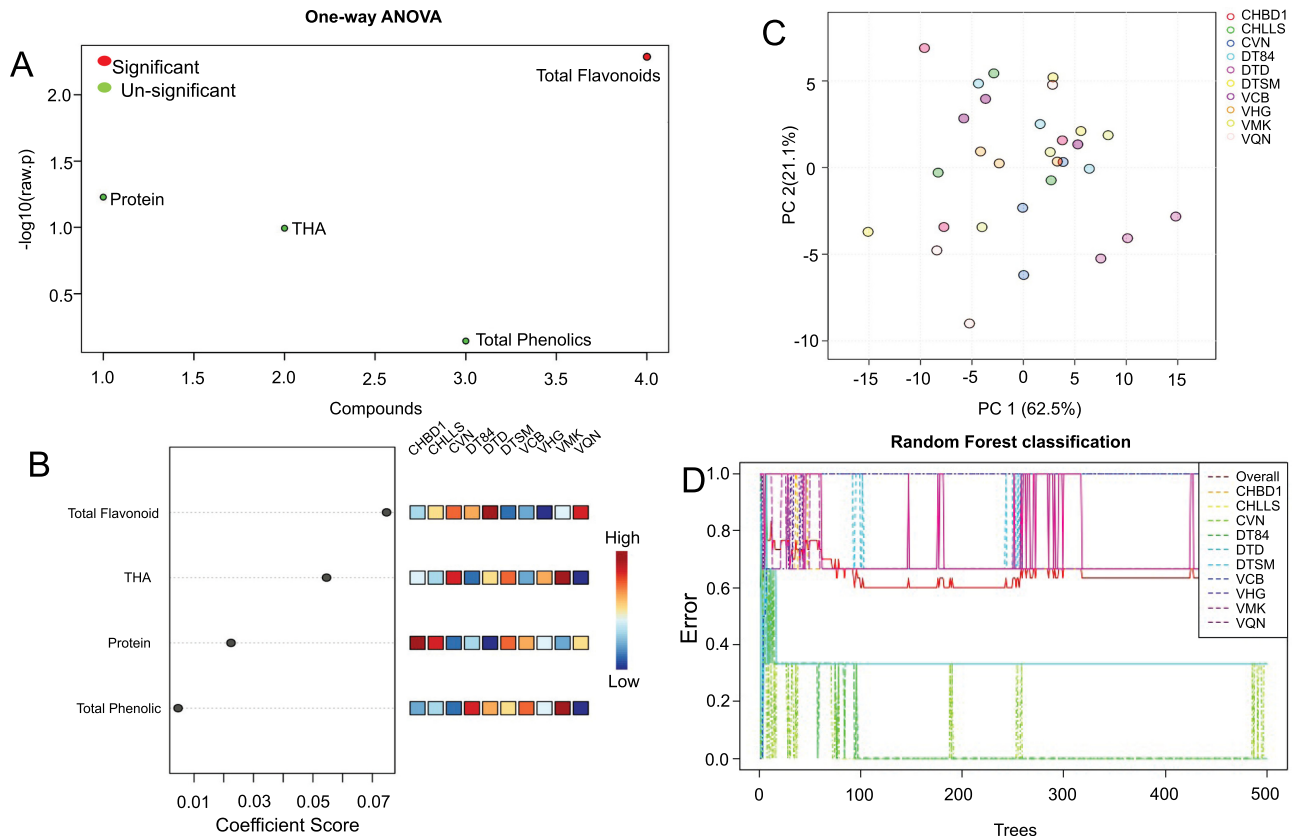


Fig. (5). Selection cultivar based on alteration of seed nutrients of soybean landrace and cultivar. **(A)** ANOVA analysis is predicting the rank of bioactive contents on the basis of metabolic profiles of soybean landraces and cultivars. The composition significant-code red color and un-significant-code green color. **(B)** Coefficient score visualization represents the relative nutrient compositions in various soybean cultivars. The levels of correlation exhibit range color from blue to red color, which is relevant low to high correlations. **(C)** Principal component analysis (PCA) on seed quality of soybean landrace and cultivars. Each point in biplot represents a single cultivar; cultivars are color-coded with the different colors. **(D)** Random Forest classification predictive model constructed using 10 cultivars for predicting the rank of cultivars based on seed nutrients concentration. Post-hoc analysis: Fisher's LSD ANOVA p-value (FDR) cutoff: 0.05.

The Pearson correlation would be used to enhance our understanding of the relationship between nutrition compositions in soybean seeds [8]. Positive correlations were detected between THA and PA or FA, while PA and FA, in most cases, correlated with the THA content. By calculating THA correlated with total PA and FA, the distinct relationships among them were observed relative to those correlated with PA and FA (Fig. S1). Excluding insignificant ones, approximately the correlation coefficients among the adjusted THA were positive. Even more intriguingly, most of the PA- and FA-corrected proteins showed a negative relationship.

3.5. Metabolites Correlation Response to Seed Weights and Seed Coat Colors

To comprehend each individual unique contribution to seed weight or seed coat color, we conducted a correlation coefficient test based on target metabolites. One of our

objectives was to determine the interrelationship between seed quality and phenotype that enhances production. In this research, every seed composition feature showed a different distribution. The seed nutrient components were found to be related to the performance of the phenotype. This study revealed a negative correlation between protein and FA concentration with a correlation coefficient of $r = -0.36^*$ ($p < 0.05$) (Table S3). Nonetheless, there is limited knowledge regarding the magnitude of this relationship. There are some correlations that may exist between seed phenotypic characteristics (seed size, seed weight, and seed color) and soybean seed nutritional characteristics.

Correlation analysis is a straightforward method for assessing the strength of the relationship on the recorded metabolite levels [2, 39]. Interpreting correlations is considered an initial step in metabolite data analysis, as the correlations arising from internal fluctuations within metabolic systems provide further insights into the

physiological condition of the seed [18]. Seed weight appeared to have a lesser impact on the sensitivity of each targeted metabolite, as there was a consistent trend of correlation responses between the metabolites, irrespective of the 100-seed dry weight, even though the significance fluctuated somewhat based on the weight (Table S3). Seed weight (100 seed dry weight) is a highly positive factor correlated to the content of protein ($r = 0.25^*$), which is contrary to the content of flavonoids in soybean seed ($r = -0.26^*$). Protein is mostly negatively correlated with the total phenolic and total flavonoid ($r = -0.30^*$ and $r = -0.36^*$, $p < 0.05$) while positively correlated with THA, irrespective of seed weight. However, the present results indicated that soybeans with a long growing period have considerably high contents of protein ($r = 0.96^{***}$, $p < 0.05$), but not for the accumulation of flavonoid ($r = -0.23^*$, $p < 0.05$) (Table S3 and Fig. S1). It is essential to highlight the correlation analysis between seed nutritional composition and genotypic (landraces and cultivar) or seed phenotypic that exist within the traits. This suggests that soybean breeders could concentrate on developing modern cultivars for specific traits of interest, a concept supported by previous studies [6, 12].

Table 2. Main interaction effects of seed coat color and seeds dry weight on the contents of from 10 soybean germplasm accessions with black (n = 1), yellowish (n = 3), and yellow (n = 6) seed coat.

Factors	Total Protein	THA	Total Phenolic	Total Flavonoid
Seed color (C)	NS	NS	*	***
Seed weight (W)	***	*	NS	NS
C X W	*	NS	NS	*

Note: NS, *, **, *** = not significant or significant at $p < 0.05$, 0.01, 0.001, respectively.

The fluctuation in seed metabolites and the variation in soybean seed nutrition compositions are strongly interrelated. Numerous studies revealed that phenotype among the properties of soybean seeds and seed coat color could be among the most significant factors to consider regarding seed quality [18, 40-42]. The variation in seed color was significantly different in soybean germplasm accessions, which indicated the total flavonoid positive correlation to seed color ($p < 0.05$) (Fig. S2). Seed color was shown to be significantly different between landraces and cultivars. Both DTD and CVN demonstrated small seed sizes, but the DTD was characterized by a black color, while CVN was yellowish, which is closely related to the exhibition of high levels of phenolic and flavonoid, respectively. These results are in agreement with the findings of a study by Choi *et al.* (2021) [42], who observed higher phenolic and flavonoid levels in soybean seeds that are black in color.

3.6. Geographical Distribution with the Seed Nutritional Characteristic

Targeted metabolites correlation response in soybean seeds may result from the influence of seed weight, as well as cultivation duration. We performed geographic analysis to determine the distribution of individual and total targeted metabolites for cultivars resources. Geographic analysis was conducted to gather comprehensive information regarding the relationship between the allocation and movement of genotypic cultivars and seed nutrients during the selection of soybean landraces and cultivars. The nutrition of soybean seeds was found to be affected by the regional distribution of cultivars cultivated in the North of Vietnam [7, 8, 43]. Various correlations may exist between geographical factors (latitude, longitude, and altitude) and the nutritional properties of soybean seeds [44]. The average nutritional compositions of seeds from all cultivars across locations were associated with geographical factors of their corresponding regions (Fig. 6). Protein content showed no significant differences between three regions, such as Hong River Delta Region (HRDR), Northwestern Region (NWR), and Northeastern Region (NTR) (Fig. 6A), which was in line with the comparison of THA content in soybean seeds of three ecoregions, HRDR, NWR, and NTR (Fig. 6B). Moreover, PA ($p < 0.05$) and FA ($p < 0.001$) content of soybean seeds in HRDR were higher compared to NWR or NTR (Fig. 6C-D).

Geographical distribution maps help to identify the region with a desirable constituent of seed components and visualize the relationship between the trend of quality characteristics and the cultivation areas. The maps also illustrate the correlation between the geographic factors and the contrast of protein and FA content in soybean seeds. In particular, provinces belonging to the Northwestern and Hong River Delta were found to be the hotspots to illustrate the contrast among the levels of protein and flavonoids in soybean seeds. Hong River delta area, especially Vinh Phuc (VP), is characterized by lowland regions classified as the third tier of elevation altitude < 500 m (comparatively high latitudinal and longitudinal coordinates). The accessions in this area showed the predominance of FA content in soybean seed composition compared to the content of protein. Conversely, the accessions in the highland area, including the Ha Giang (HG) plateaus belonging to the second altitude tier ($1000 \text{ m} < \text{altitude} < 2000 \text{ m}$), which are also at relatively decreased latitudinal and longitudinal coordinates, exhibited a contrasting profile of seed nutrition composition with higher content of protein comparing to FA level (Fig. 7). Correlation values among the seed nutrition components were also substantial and positive for the three cultivation regions. These results provide new insight into the correlation of the variability of flavonoids in soybean seeds with soybean germplasm and various geographical factors. Overall, our findings indicate that HRDR accessions may be more suitable for producing high levels of flavonoids in black soybeans, such as the DTD cultivar.

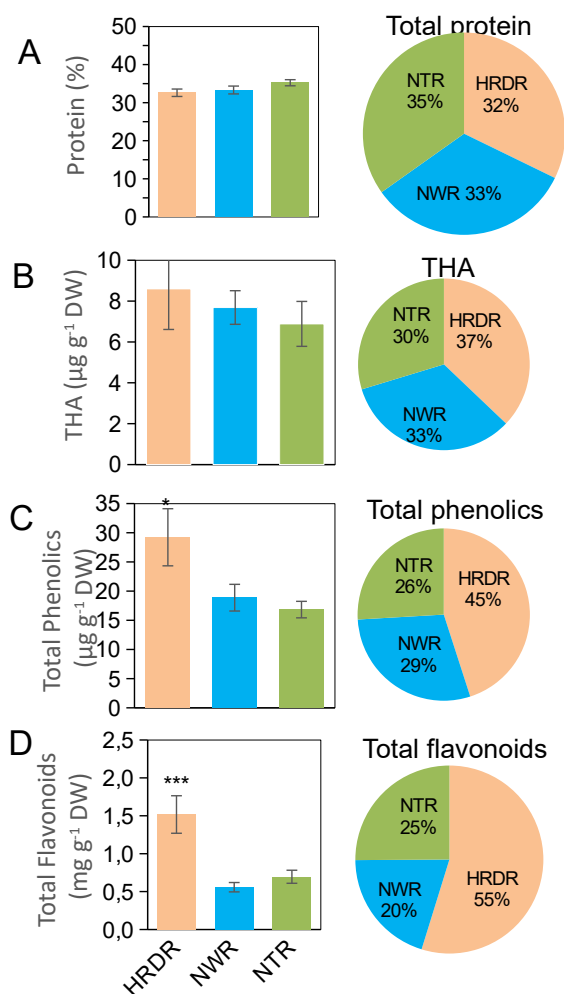


Fig. (6). Distribution of seed quality in soybean cultivars collected from three ecoregions. **(A)** Protein, **(B)** Hydroxycinnamic acid (THA). **(C)** Total phenolic acid, **(D)** Total flavonoid in soybean seed collected from the three regions Hong River Delta Region (HRDR), Northwestern Region (NWR), and Northeastern Region (NTR). The data represents mean \pm SE (n = 3). Asterisks indicate significant differences between the different ecoregions as determined by Duncan's t-test; * $p < 0.05$, * * * $p < 0.001$.

4. DISCUSSION

Developing soybean cultivars with high seed nutrients is one of the most challenging strategies for soybean improvements. It has been widely recognized that the biochemical response of designated metabolites is influenced by both genetic factors and environmental fluctuations. Although there are many studies on metabolic variations in soybean seeds, data regarding the responses of protein and flavonoids to a variety of colors of seed coats and the dry weight of seeds remains limited. The current data demonstrated that the fluctuations in metabolism were affected by means of both seed coat coloration and dry weight of 100 seeds (Fig. 1 and Table 1). The phenotypic responses to seed weight and seed coat color were found to play a role in mediating the varying responses linked to metabolic categorization (Fig. 4) and associations (Fig. S2). Furthermore, these individual metabolites were highly interconnected, regardless of differences in seed weight or seed coat color (Table S3

and Fig. 5). As per our knowledge, this study is the first to present the inverse relationship between protein and flavonoid content concerning seed coat color across a broad spectrum of primary targeted metabolites, in addition to the dry weight of 100 seeds in soybean seed germplasm accessions. The dynamic variability of protein and flavonoid contents observed in this research and earlier research is probably due to the various groups of accessions and the extensive cultivar collections present in this research that aid in selecting genotypes with improved levels of bioactive compounds for breeding. It can be concluded that soybean cultivars with relatively high protein or flavonoid content have the potential to be used for food products, including VHG or DTD.

Soybean quality is influenced not only by the total protein concentration but also by the profiles of phenolic and flavonoid compounds. It has been known that hydroxycinnamic acid (THA), a cinnamate compound, is involved in the synthesis of phenolic compounds and

flavonoids (Fig. 3A) [45]. Thus, the concentrations of seed hydroxycinnamic acid and phenolic compounds or flavonoids are inherently strongly linked due to THA consisting of phenolic compounds and flavonoids. The measurement methods commonly used in previous studies were based on seed weight; thus, they could not illustrate the impact of THA on the content of phenolics and flavonoids in specific genotype variants [18, 46]. In the present study, the relationship between the content of phenolics and flavonoids and THA was investigated simultaneously with dry weight-based measurement. We aimed to determine if phenolic and flavonoid contents in seeds correlated with THA in all cultivars. The result illustrated that the THA content was correlated to both phenolic and flavonoid content in various soybean cultivars, especially in DTD and CVN. A major impact of THA was found on the composition of phenolics and flavonoids, indicating a dependent relationship between THA and phenolic compounds. However, the majority of THA simultaneously accumulated in VMK, VHJ, CVN, and

DTSM, but not phenolics and flavonoids (Fig. 3B-D). Therefore, the impact of THA on phenolic or flavonoid contents could be eliminated by discovering genotypic variants specific to the THA profile. These data suggested that a rise in THA may promote the amount of PA and FA in terms of definite content but not necessarily affect the content of THA. It has been observed that since THA is not strongly associated with the modified content of PA and FA, the relationship is more complex. The variation in cultivars leads to a certain proportion of THA, and the modification of soybean phenolic content might create a wide-ranging impact on the seed nutrient profile. In some cases, the positive genetic correlation of THA and phenolic or flavonoid became negative after modification of THA content, while in other cases, the correlation remained positive (Fig. S1). This implied different cultivars for THA with and without phenolic-based adjustment in various soybeans. It additionally implied that the THA profile might be optimized without modifying the overall amount of phenolic and total flavonoid composition, which could paradoxically influence soybean color or seed yield.

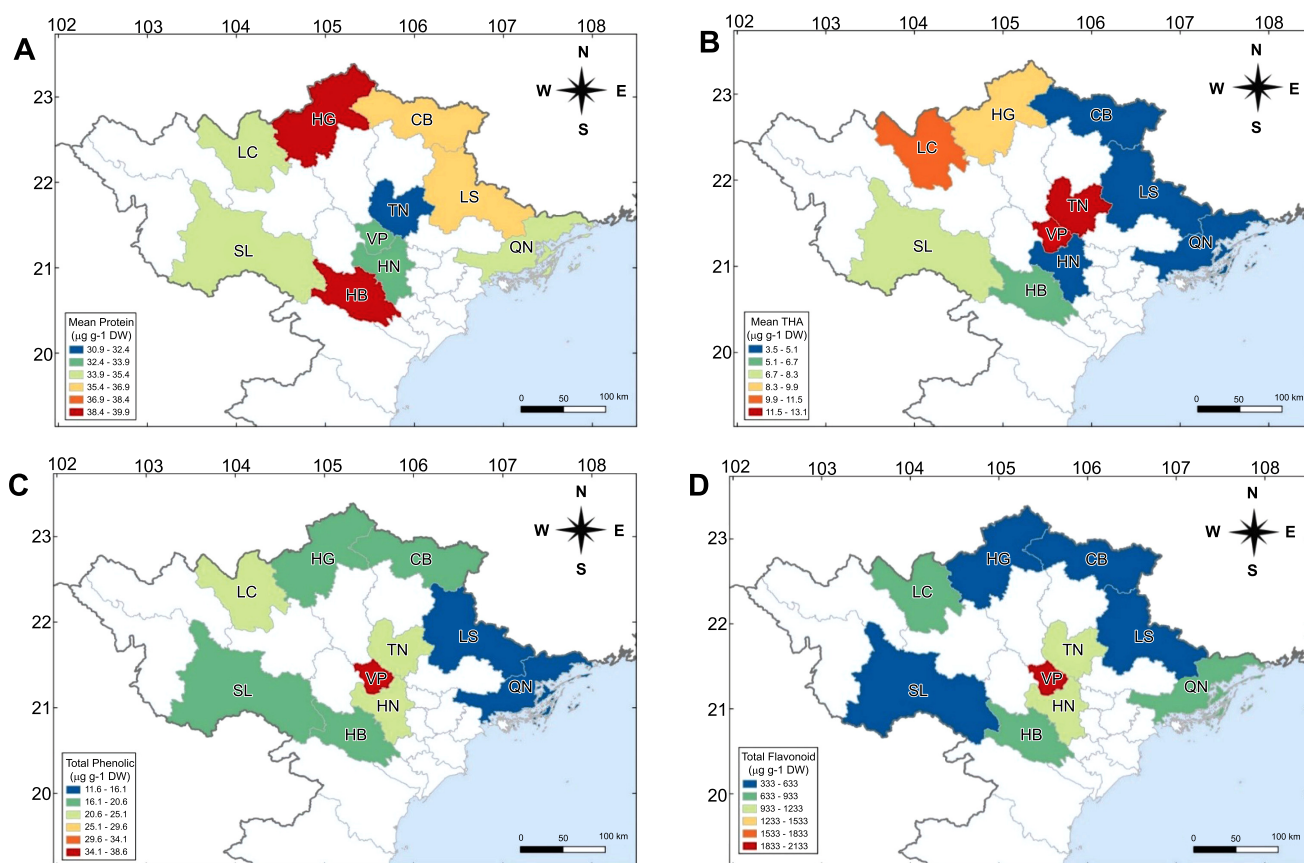


Fig. (7). Geographical distribution of seed nutrition concentration in soybean seeds, mapped according to the region of accession origin. (A) Total protein, (B) Total hydroxycinnamic acid (THA), (C) Total phenolic acid, (D) Total flavonoid. The ten provinces are represented by three ecoregions Hong River Delta (HRDR), Northeastern region (NTR); and Northwestern region (NWR). Soybean genotypes collected from Ha Noi (HN), Vinh Phuc (VP), Lao Cai (LC), Cao Bang (CB), Ha Giang (HG), Quang Ninh (QN), Hoa Binh (HB), Lang Son (LS), Thai Nguyen (TN), Son La (SL). Geographical distribution maps of soybean seed nutrition were conducted with QGIS 10.0 (<https://www.qgis.org/en/site/>) using ordinary interpolation.

Seed weight is influenced by biochemical responses, which can subsequently impact the metabolite contents, yield, and quality of soybean seeds. Many studies have reported on variations in protein, phenolic, and flavonoid, along with their relationship with genetic factors or environmental factors, while investigating the impacts of both elements on the yield and quality of soybean seeds [2, 14, 18, 47]. Recently, Zhang *et al.* (2018) [12] illustrated that the amino acid concentration associated with seed weight and total protein exhibited a distinct genetic basis. Furthermore, 100-seed weight was influenced by the levels of sugars, protein, oil, phenolic, and fatty acids [18, 41]. The dry mass of 100 seeds was likely greater in yellow soybean seed germplasm lines compared to those of other seed coat colors (Fig. 2). The difference in dry weight for 100 seeds between black and yellow soybean accessions ranged from 8 to 15 grams even though both colored soybean accessions were widely distributed in terms of their 100-seed weight. The dry weight of 100 seeds affected the selected metabolites differently based on seed coat color. In addition, protein content showed a positive correlation with seed weight, while flavonoid content was positively associated with the dry weight of 100 seeds. The THA, phenolic, and flavonoid levels from yellow and black soybean accessions were inconsistently associated with seed weight, while the protein levels in yellow soybean accessions demonstrated a consistently significant relationship with seed weight. This result suggests that protein levels would remain consistently steady within a soybean lineage, regardless of variations in the dry weight of 100 seeds. In other studies, protein levels were found to increase positively with the dry weight of 100 seeds in black soybean seeds, but this was not observed in yellow soybean seeds [6, 19, 48]. In yellow seeds, the levels of total protein were correlated with 100-seed weight. In another study, Xu *et al.* (2022) [19] reported that protein compounds are affected by seed weight in *Glycine max* (L.). Furthermore, large soybean seeds exhibited significantly higher levels of starch, sucrose, protein, fatty acid, and phenolic than small seeds [18-20]. It has also been indicated that seed size is positively associated with protein and sugar content, while it is negatively correlated with isoflavones. In the case of yellow soybean accessions, only protein progressively increased with seed weight; however, the amount of flavonoid decreased. In soybeans, a wide range of protein contents was found in a large seed compared to a small seed [17, 19]. Although the metabolic variation was notably distinct in the ANOVA clustering pattern (Fig. 4A, C), the correlation coefficient exhibited a high total protein contribution to the dry weight of 100 seeds (Table 2 and S3). Moreover, a previous study suggested a significant promise for enhancing soybean seed nutrients without compromising yield because protein has a positive correlation with total sugar and sucrose [18, 25, 27]. Furthermore, considering that the soybean germplasm accessions in this study were sourced from various maturity groups, it was intriguing to investigate whether the protein, phenolic, and flavonoid in seeds were correlated with days to maturity (time of cultivation). The results indicated a significant and strong

connection between protein concentration and days until maturity, with a correlation coefficient of 0.96 ($P = 0.85$). At the same time, the association between phenolic or flavonoid levels and days to maturity was observed to be quite weak, with a correlation coefficient of -0.29 and -0.22 ($P = 0.90$; $P = 0.82$) (Fig. S1). This suggested that protein content in seeds might be genotypically influenced by days to maturity. The rise in concentrations of these targeted metabolites was closely associated with the elevation in seed dry weight as the seeds progressed through the growth and development stages [49].

Besides the effect of seed weight on the biochemical response of focused metabolites, the color of the seed coats also affected the level of protein, THA, phenolic, and flavonoid. In the present study, we found that fluctuations in protein and flavonoid levels had a major effect on seed coat color. The impact of seed coat color did not appear to be significant on protein levels in locally collected soybean accessions, whereas notable variation was observed among the collection locations of these accessions (Fig. 4). This result is in agreement with the findings of a study by Kim *et al.* (2007) [24], who reported that the color of the seed coat did not influence the negative correlation of protein and lipid contents. In contrast, the nutritional compositions in seeds with higher phenolic and flavonoid content were found in black seed coat color, as detected in DTD, which hinted that the phenolic and flavonoid levels were highly affected by seed coat color. The influence of seed coat color did not seem to significantly affect protein levels in soybean accessions gathered from local sources. Notable differences were observed across various collection locations of these accessions [23, 47]. Nonetheless, in the present study, the total phenolic and flavonoid content was higher in the soybean seeds with black coat color compared to those with yellow seed coat color. The coloring of black soybean seeds was linked to the function of a UDP-glucose flavonoid-3-O-glycosyltransferase involved in anthocyanin synthesis compared to the activity observed in yellow soybean seeds. Interestingly, the color of black soybean seeds was strongly correlated with isoflavone, a flavonoid compound, but comparatively lower in anthocyanin levels throughout the seed maturation phase [23]. The dynamic of the change of targeted metabolites in this study was significantly interrelated with the collected soybean accession and was influenced by seed coat color. Further, an increase in the amount of either protein or bioactive compounds (phenolic and flavonoid) beyond these allowable ranges would dramatically inhibit the other component. Owing to the low protein content in DTD and CVN cultivars, it is expected that their flavonoid levels will increase as a result of a negative correlation (Fig. 4B). In other words, it is impossible to simultaneously increase protein, phenolic, and flavonoid levels within a certain range. To the best of our understanding, this is the first research to uncover the influence of seed coat color on a wide array of dynamics related to protein and flavonoids and their negative association, along with soybean seed germplasm accessions. The correlation between protein

and phenolic or flavonoid is regarded as an assessment of metabolite interaction caused by the correlations induced by internal fluctuation of the metabolic system [6, 8]. Compared to other traits, flavonoids exhibited a broader distribution in seed coat color when compared to landraces and cultivars. Flavonoid was identified as the comparatively most stable component when compared with the variation coefficient values of others. It implied that flavonoids have strong genotypic control, and the high flavonoid genotypes are likely to maintain performance in the DTD soybean cultivar. The variation in protein, phenolic, and flavonoid levels in soybean seeds can be linked to the influence of genotype traits [6, 9, 44], along with environmental factors [10, 11].

Seed nutrients exhibited significant variations across different ecoregions of origin, suggesting that geography may play an important role in the variety of soybeans in Vietnam. These results indicated that HRDR accessions are likely to have the highest concentrations of flavonoid components, linked to the increased genetic diversity observed in NTR and NWR accessions (Fig. 5). The geographic distribution map of protein and THA levels indicated a declining trend as one moves southward (toward lower latitudes). However, total phenolic and flavonoid contents were found to be highly accumulated in the Vinh Phuc (VP), a filed flat land of the HRDR area (low latitude). Obviously, geographical differentiation conditions could contribute substantially to the genetic differentiation of soybeans, resulting in variations in seed quality traits [9, 50, 51]. Crucially, this finding may support breeders in broadening the range of germplasm and encourage the application of these valuable genetic resources in breeding initiatives centered on protein or flavonoid content. In light of our findings, quality traits of soybean seeds, such as amino acid and protein, fatty acid, flavonoid, isoflavone, anthocyanins, and tocopherols, are influenced by the ecoregion of origin [6-8, 12, 52, 53]. Overall, the significant differences in protein, phenolic, and flavonoid compositions across ecoregions are associated with environmental changes, geographical positions, and growing seasons [8, 10, 12, 54]. Thus, it is essential to highlight that nutritional composition varies significantly with the geographical origin of soybean accessions, implying that soybean breeders ought to prioritize the origin of these accessions when developing contemporary cultivars for a specific trait of interest, and this notion is supported by earlier studies [12]. This variation enables breeders to select highly adapted candidate accessions with increased protein or flavonoid content from different locations, thereby contributing to improvements in soybean quality breeding.

CONCLUSION

This study is significant for determining the qualitative soybean cultivars based on the intricate relationships among seed constituents, especially the negative correlation among protein, phenolic compounds, and flavonoid levels. It reveals the genotype basis of quality alterations in soybean landraces and cultivars. Additionally, this study provides insight into how genetic

and geographical factors govern the seed compounds. Such variations in nutritional content among soybean cultivars can be leveraged for significant applications in the food and pharmaceutical industries. Further validation of the causal relationship between nutrition variation and the phenotypic effect linked to soybean quality, as well as understanding their relationship with yield, will be the focus of the next study.

AUTHORS' CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: data collection: V.H.H., V.T.N., T.T.L., T.P.A.D., D.V.D, X.B.N., T.D.N., T.D.T.; validation: M.N.; draft manuscript: V.H.L., A.T.T., D.H.T. All authors reviewed the results and approved the final version of the manuscript.

LIST OF ABBREVIATIONS

TUAF	= Thai Nguyen University of Agriculture and Forestry
PRC	= Plant Resource Center
PIs	= Plant Introductions
RP-HPLC	= Reverse Phase High-Performance Liquid Chromatography
ANOVA	= Analysis of Variance
PCA	= Principal Component Analysis

RESEARCH INVOLVING PLANTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data are available with the links provided in the manuscript. Supplementary data to this article are included in this published article and can be found online.

FUNDING

This work was supported by national funding from the Ministry of Education and Training, Vietnam, which supported the funding acquisition (grant code number B2022-TNA-42).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The authors would like to thank the Plant Resource Center (PRC) for providing soybean germplasm accessions (It also supports Anh Phuong Thi Dang's experiment, a student funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code [VINIF. 2024.TS.045]).

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

- [1] Xu B, Chang SKC. Antioxidant capacity of seed coat, dehulled bean, and whole black soybeans in relation to their distributions of total phenolics, phenolic acids, anthocyanins, and isoflavones. *J Agric Food Chem* 2008; 56(18): 8365-73.
<http://dx.doi.org/10.1021/jf801196d> PMID: 18729453
- [2] Lin H, Rao J, Shi J, *et al.* Seed metabolomic study reveals significant metabolite variations and correlations among different soybean cultivars. *J Integr Plant Biol* 2014; 56(9): 826-36.
<http://dx.doi.org/10.1111/jipb.12228> PMID: 24942044
- [3] Singh BP, Yadav D, Vij S. Soybean bioactive molecules: Current trend and future prospective Bioactive Molecules in Food. Cham: Springer International Publishing 2019; pp. 267-94.
http://dx.doi.org/10.1007/978-3-319-78030-6_4
- [4] Xu JL, Shin JS, Park SK, *et al.* Differences in the metabolic profiles and antioxidant activities of wild and cultivated black soybeans evaluated by correlation analysis. *Food Res Int* 2017; 100(Pt 2): 166-74.
<http://dx.doi.org/10.1016/j.foodres.2017.08.026> PMID: 28888437
- [5] Zhang RF, Zhang FX, Zhang MW, *et al.* Phenolic composition and antioxidant activity in seed coats of 60 Chinese black soybean (*Glycine max* L. Merr.) varieties. *J Agric Food Chem* 2011; 59(11): 5935-44.
<http://dx.doi.org/10.1021/jf201593n> PMID: 21548651
- [6] Hyeon H, Xu JL, Kim JK, Choi Y. Comparative metabolic profiling of cultivated and wild black soybeans reveals distinct metabolic alterations associated with their domestication. *Food Res Int* 2020; 134: 109290.
<http://dx.doi.org/10.1016/j.foodres.2020.109290> PMID: 32517920
- [7] Azam M, Zhang S, Qi J, *et al.* Profiling and associations of seed nutritional characteristics in Chinese and USA soybean cultivars. *J Food Compos Anal* 2021; 98: 103803.
<http://dx.doi.org/10.1016/j.jfca.2021.103803>
- [8] Abdelghany AM, Zhang S, Azam M, *et al.* Profiling of seed fatty acid composition in 1025 Chinese soybean accessions from diverse ecoregions. *Crop J* 2020; 8(4): 635-44.
<http://dx.doi.org/10.1016/j.cj.2019.11.002>
- [9] Xiong M, Zhao M, Lu ZX, Balasubramanian P. Genotypic variation for phenolic compounds in developing and whole seeds, and storage conditions influence visual seed quality of yellow dry bean genotypes. *Can J Plant Sci* 2020; 100(3): 284-95.
<http://dx.doi.org/10.1139/cjps-2019-0153>
- [10] La VH, Tran DH, Han VC, *et al.* Drought stress-responsive abscisic acid and salicylic acid crosstalk with the phenylpropanoid pathway in soybean seeds. *Physiol Plant* 2023; 175(5): e14050.
<http://dx.doi.org/10.1111/pp1.14050> PMID: 37882260
- [11] Veremeichik GN, Grigorchuk VP, Butovets ES, *et al.* Isoflavonoid biosynthesis in cultivated and wild soybeans grown in the field under adverse climate conditions. *Food Chem* 2021; 342: 128292.
<http://dx.doi.org/10.1016/j.foodchem.2020.128292> PMID: 33069538
- [12] Zhang J, Wang X, Lu Y, *et al.* Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. *Mol Plant* 2018; 11(3): 460-72.
<http://dx.doi.org/10.1016/j.molp.2017.12.016> PMID: 29305230
- [13] Riedl KM, Lee JH, Renita M, St Martin SK, Schwartz SJ, Vodovotz Y. Isoflavone profiles, phenol content, and antioxidant activity of soybean seeds as influenced by cultivar and growing location in Ohio. *J Sci Food Agric* 2007; 87(7): 1197-206.
<http://dx.doi.org/10.1002/jsfa.2795>
- [14] Li X, Kamala S, Tian R, *et al.* Identification and validation of quantitative trait loci controlling seed isoflavone content across multiple environments and backgrounds in soybean. *Mol Breed* 2018; 38(1): 8.
<http://dx.doi.org/10.1007/s11032-017-0768-8>
- [15] Wu D, Li D, Zhao X, *et al.* Identification of a candidate gene associated with isoflavone content in soybean seeds using genome-wide association and linkage mapping. *Plant J* 2020; 104(4): 950-63.
<http://dx.doi.org/10.1111/tpj.14972> PMID: 32862479
- [16] Meng N, Yu BJ, Guo JS. Ameliorative effects of inoculation with *Bradyrhizobium japonicum* on *Glycine max* and *Glycine soja* seedlings under salt stress. *Plant Growth Regul* 2016; 80(2): 137-47.
<http://dx.doi.org/10.1007/s10725-016-0150-6>
- [17] Kim SL, Berhow MA, Kim JT, Chi HY, Lee SJ, Chung IM. Evaluation of soyasaponin, isoflavone, protein, lipid, and free sugar accumulation in developing soybean seeds. *J Agric Food Chem* 2006; 54(26): 10003-10.
<http://dx.doi.org/10.1021/jf062275p> PMID: 17177534
- [18] Lee J, Hwang YS, Kim ST, *et al.* Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. *Food Chem* 2017; 214: 248-58.
<http://dx.doi.org/10.1016/j.foodchem.2016.07.066> PMID: 27507473
- [19] Xu C, Wu T, Yuan S, *et al.* Can soybean cultivars with larger seed size produce more protein, lipids, and seed yield? A Meta-Analysis. *Foods* 2022; 11(24): 4059.
<http://dx.doi.org/10.3390/foods11244059> PMID: 36553799
- [20] Duan Z, Li Q, Wang H, He X, Zhang M. Genetic regulatory networks of soybean seed size, oil and protein contents. *Front Plant Sci* 2023; 14: 1160418.
<http://dx.doi.org/10.3389/fpls.2023.1160418> PMID: 36959925
- [21] Kumar V, Rani A, Solanki S, Hussain SM. Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. *J Food Compos Anal* 2006; 19(2-3): 188-95.
<http://dx.doi.org/10.1016/j.jfca.2005.06.005>
- [22] Cho KM, Ha TJ, Lee YB, *et al.* Soluble phenolics and antioxidant properties of soybean (*Glycine max* L.) cultivars with varying seed coat colours. *J Funct Foods* 2013; 5(3): 1065-76.
<http://dx.doi.org/10.1016/j.jff.2013.03.002>
- [23] Choi YM, Yoon H, Lee S, *et al.* Comparison of isoflavone composition and content in seeds of soybean (*Glycine max* (L.) Merrill) germplasms with different seed coat colors and days to maturity. *Korean J Plant Res* 2020; 33: 558-77.
- [24] Kim SL, Lee YH, Chi HY, Lee SJ, Kim SJ. Diversity in lipid contents and fatty acid composition of soybean seeds cultivated in Korea. *Korean J Crop Sci* 2007; 52(3): 348-57.
- [25] Lee HS, Son BY. Variation of sugar content and its relationship with some major characteristics in collection of colored soybean. *Hangug Jagmul Haghoeji* 1993; 37: 521-7.
- [26] Kim SL, Chi HY, Son JR, Park NK, Ryu SN. Physicochemical characteristics of soybean seed coat and their relationship to seed lustre. *Hangug Jagmul Haghoeji* 2005; 50: 123-31.
- [27] Wilcox JR, Shibles RM. Interrelationships among seed quality attributes in soybean. *Crop Sci* 2001; 41(1): 11-4.
<http://dx.doi.org/10.2135/cropsci2001.41111x>
- [28] Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 2012; 9(7): 671-5.
<http://dx.doi.org/10.1038/nmeth.2089> PMID: 22930834
- [29] Kaur C, Kapoor HC. Anti-oxidant activity and total phenolic content of some Asian vegetables. *Int J Food Sci Technol* 2002; 37(2): 153-61.
<http://dx.doi.org/10.1046/j.1365-2621.2002.00552.x>
- [30] Štefan MB, Vuković Rodríguez J, Blažeković B, Kindl M, Vladimír-Knežević S. Total hydroxycinnamic acids assay: Prevalidation and application on *Lamiaceae* species. *Food Anal Methods* 2014; 7(2): 326-36.
<http://dx.doi.org/10.1007/s12161-013-9630-8>
- [31] Zhishen J, Mengcheng T, Jianming W. The determination of flavonoid contents in mulberry and their scavenging effects on superoxide radicals. *Food Chem* 1999; 64(4): 555-9.
[http://dx.doi.org/10.1016/S0308-8146\(98\)00102-2](http://dx.doi.org/10.1016/S0308-8146(98)00102-2)
- [32] Das PR, Eun JB. A comparative study of ultra-sonication and agitation extraction techniques on bioactive metabolites of green tea extract. *Food Chem* 2018; 253: 22-9.
<http://dx.doi.org/10.1016/j.foodchem.2018.01.080> PMID: 29305230

- 29502824
- [33] Oñate-Sánchez L, Vicente-Carbajosa J. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. BMC Res Notes 2008; 1(1): 93. <http://dx.doi.org/10.1186/1756-0500-1-93> PMID: 18937828
- [34] Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{-ΔΔCT} Method. Methods 2001; 25(4): 402-8. <http://dx.doi.org/10.1006/meth.2001.1262> PMID: 11846609
- [35] Radočaj D, Jurišić M, Gašparović M, Plaščak I. Optimal soybean (*Glycine max* L.) Land suitability using GIS-based multicriteria analysis and sentinel-2 multitemporal images. Remote Sens (Basel) 2020; 12(9): 1463. <http://dx.doi.org/10.3390/rs12091463>
- [36] Meinke DW, Chen J, Beachy RN. Expression of storage-protein genes during soybean seed development. Planta 1981; 153(2): 130-9. <http://dx.doi.org/10.1007/BF00384094> PMID: 24276763
- [37] Ha DH, Kim HJ. Mutation of storage protein gene using CRISPR/Cas9 removed α'-subunit of β-conglycinin in soybean seeds. Plant Biotechnol Rep 2023; 17(6): 939-45. <http://dx.doi.org/10.1007/s11816-023-00880-3>
- [38] Dhaubhadel S, Gijzen M, Moy P, Farhangkhoe M. Transcriptome analysis reveals a critical role of *CHS7* and *CHS8* genes for isoflavonoid synthesis in soybean seeds. Plant Physiol 2007; 143(1): 326-38. <http://dx.doi.org/10.1104/pp.106.086306> PMID: 17098860
- [39] Park SY, Lim SH, Ha SH, et al. Metabolite profiling approach reveals the interface of primary and secondary metabolism in colored cauliflowers (*Brassica oleracea* L. ssp. botrytis). J Agric Food Chem 2013; 61(28): 6999-7007. <http://dx.doi.org/10.1021/jf401330e> PMID: 23782237
- [40] Lim YJ, Kwon SJ, Qu S, Kim DG, Eom SH. Antioxidant contributors in seed, seed coat, and cotyledon of γ-ray-induced soybean mutant lines with different seed coat colors. Antioxidants 2021; 10(3): 353. <http://dx.doi.org/10.3390/antiox10030353> PMID: 33652948
- [41] Choi YM, Yoon H, Lee S, et al. Isoflavones, anthocyanins, phenolic content, and antioxidant activities of black soybeans (*Glycine max* (L.) Merrill) as affected by seed weight. Sci Rep 2020; 10(1): 19960. <http://dx.doi.org/10.1038/s41598-020-76985-4> PMID: 33203918
- [42] Choi YM, Yoon H, Shin MJ, et al. Metabolite contents and antioxidant activities of soybean (*Glycine max* (L.) Merrill) seeds of different seed coat colors. Antioxidants 2021; 10(8): 1210. <http://dx.doi.org/10.3390/antiox10081210> PMID: 34439461
- [43] Gebregziabher BS, Zhang S, Azam M, et al. Natural variations and geographical distributions of seed carotenoids and chlorophylls in 1 167 Chinese soybean accessions. J Integr Agric 2023; 22(9): 2632-47. <http://dx.doi.org/10.1016/j.jia.2022.10.011>
- [44] Chen Q, Wang X, Yuan X, et al. Comparison of phenolic and flavonoid compound profiles and antioxidant and α-glucosidase inhibition properties of cultivated soybean (*Glycine max*) and wild soybean (*Glycine soja*). Plants 2021; 10(4): 813. <http://dx.doi.org/10.3390/plants10040813> PMID: 33924154
- [45] Dong NQ, Lin HX. Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. J Integr Plant Biol 2021; 63(1): 180-209. <http://dx.doi.org/10.1111/jipb.13054> PMID: 33325112
- [46] Arai S, Suzuki H, Fujimaki M, Sakurai Y. Studies on flavor components in soybean. Part II. Phenolic acid in defatted soybean flour. Agric Biol Chem 1966; 30(4): 364-9. <http://dx.doi.org/10.1080/00021369.1966.10858609>
- [47] Desta KT, Hur OS, Lee S, et al. Origin and seed coat color differently affect the concentrations of metabolites and antioxidant activities in soybean (*Glycine max* (L.) Merrill) seeds. Food Chem 2022; 381: 132249. <http://dx.doi.org/10.1016/j.foodchem.2022.132249> PMID: 35114623
- [48] Kim JK, Park SY, Lim SH, Yeo Y, Cho HS, Ha SH. Comparative metabolic profiling of pigmented rice (*Oryza sativa* L.) cultivars reveals primary metabolites are correlated with secondary metabolites. J Cereal Sci 2013; 57(1): 14-20. <http://dx.doi.org/10.1016/j.jcs.2012.09.012>
- [49] Kambhampati S, Aznar-Moreno JA, Bailey SR, et al. Temporal changes in metabolism late in seed development affect biomass composition. Plant Physiol 2021; 186(2): 874-90. <http://dx.doi.org/10.1093/plphys/kiab116> PMID: 33693938
- [50] Lee JH, Choung MG. Comparison of nutritional components in soybean varieties with different geographical origins. J Korean Soc Appl Biol Chem 2011; 54(2): 254-63. <http://dx.doi.org/10.3839/jksabc.2011.040>
- [51] Zhou Z, Jiang Y, Wang Z, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 2015; 33(4): 408-14. <http://dx.doi.org/10.1038/nbt.3096> PMID: 25643055
- [52] Kim EH, Lee OK, Kim JK, et al. Isoflavones and anthocyanins analysis in soybean (*Glycine max* (L.) Merrill) from three different planting locations in Korea. Field Crops Res 2014; 156: 76-83. <http://dx.doi.org/10.1016/j.fcr.2013.10.020>
- [53] Ghosh S, Zhang S, Azam M, et al. Seed tocopherol assessment and geographical distribution of 1151 Chinese soybean accessions from diverse ecoregions. J Food Compos Anal 2021; 100: 103932. <http://dx.doi.org/10.1016/j.jfca.2021.103932>
- [54] Song W, Yang R, Yang X, et al. Spatial differences in soybean bioactive components across China and their influence by weather factors. Crop J 2018; 6(6): 659-68. <http://dx.doi.org/10.1016/j.cj.2018.05.001>