Research Square

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

Targeted Profiling and Interpretable Machine Learning Reveal Crop-Use-Associated Glucosinolate Signatures in Brassica juncea

Sreeparna Chowdhury
National Institute of Agrobiodiversity Center, National Institute of Agricultural Sciences, Rural Development Administration

Kanivalan Iwar
National Institute of Agrobiodiversity Center, National Institute of Agricultural Sciences, Rural Development Administration

Seong-Hoon Kim
shkim0819@korea.kr

National Institute of Agrobiodiversity Center, National Institute of Agricultural Sciences, Rural Development Administration

Research Article

Keywords: Machine learning, Brassica juncea, Glucosinolate, Crop-use forms, Mustard types

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11359378/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background Brassica juncea

has morphologically and agronomically diverse cultivated forms used for vegetative, root, and seed production and consumption. Although glucosinolate (GSL) composition varies substantially within *B. juncea*, the extent to which this crop-use diversification is reflected in leaf GSL composition remains underexplored. This study aimed to characterize leaf GSL variation among accessions representing vegetative (VM), root (RM), and seed mustard (SM) types, and to identify GSL features contributing to crop-use type discrimination.

Results

The results showed significant differences at both the crop-use forms and accession levels in *B. juncea*. Aliphatic GSL accounted for more than 98% of the total GSL in all three mustard-types, with sinigrin as the predominant compound. RM accessions exhibited lower total GSL than VM and SM, but comparatively higher 4-hydroxyglucobrassicin and neoglucobrassicin. PCA positioned RM in a distinct region, whereas SM overlapped substantially with the broader VM distribution. Pairwise machine learning (ML) further supported this pattern, with linear support vector machine and random forest models discriminating RM from VM and SM, respectively, while VM and SM exhibited weaker separation. Shapley additive explanations (SHAP) analysis detected recurrent contributions of 4-hydroxyglucobrassicin, neoglucobrassicin, sinigrin, glucocheirolin, and related GSLs to RM-associated differentiation.

Conclusion

The findings establish a biochemical framework for *B. juncea* germplasm characterization and support the prioritization of contrasting genetic resources and GSL signatures for future breeding and mustard improvement.

Background

Brassica juncea (L.) Czern. & Coss., commonly known as Indian mustard, is an economically important multipurpose crop cultivated as an oilseed, condiment, and vegetable, with different forms utilized for their seeds, leaves, stems, or roots. The crop is valued for its nutritional constituents, *viz.* vitamins, minerals, carotenoids, glucosinolates, and phenolic compounds [1, 2]. It is an allotetraploid crop species (AABB, 2n = 36) derived from the diploid progenitors *B. rapa* (AA) and *B. nigra* (BB), with extensive morphological and agronomic diversity. This diversification has given rise to distinct cultivated forms differing markedly in their principal harvested organs and patterns of use. Accordingly, based on crop use and morphology, four major subspecies have been proposed, including subsp. *juncea* (seed

mustard), subsp. *integrifolia* (leaf mustard), subsp. *napiformis* (root mustard), and subsp. *tumida* (stem mustard), with additional varietal forms occurring within these broad groups [3, 4]. Such diversity provides a biologically relevant context for examining biochemical differentiation among cultivated mustard forms.

GSLs are specialized metabolites of Brassicaceae family, and their composition contributes to both plant defense mechanism as well as characteristic pungency and quality attributes of mustard crops [5]. GSL accumulation in *B. juncea* is strongly influenced by genetic background and tissue-specific regulation. Functional analysis of GSL transporters (GTR1 and GTR2) has demonstrated distinct regulation and distribution of these metabolites among vegetative and reproductive tissues [6], while substantial quantitative and compositional variation has been documented among *B. juncea* germplasm across plant organs and developmental stages [7]. Despite this reported variation, comparative information on leaf GSL composition across contrasting cultivated forms of *B. juncea* remains limited.

Previous comparative studies on root, stem, and leaf mustard have highlighted distinct GSL and related phytochemical profiles in seeds and sprouts [8, 9]. However, studies of leaf GSL composition have predominantly focused on leaf-mustard germplasm [10, 11]. Consequently, whether contrasting cultivated forms of *B. juncea* retain distinguishable GSL profiles when evaluated in a common leaf tissue remains relatively underexplored. Determining whether such crop-use diversification is accompanied by consistent differences in leaf GSL composition is therefore important for characterizing germplasm for identifying contrasting genetic resources.

Assessing such biochemical variation requires analytical approaches capable of capturing patterns across multiple metabolites. Multivariate analysis is particularly useful for resolving overall compositional structure and group-associated patterns that may not be evident when GSL compounds are considered individually. For instance, principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) can summarize major sources of variation and identify variables associated with group segregation in metabolomic datasets [12]. The discriminatory potential of GSL profiles has been demonstrated across Brassicaceae crops, where hierarchical clustering analysis (HCA) and PCA of quantitative GSL composition distinguished 12 vegetable types and broadly reflected their morphological classification highlighting taxonomic and phenotypic patterns [13]. Similarly, in Chinese cabbage germplasm, multivariate GSL profiling resolved groups associated with germplasm status and geographical origin with PCA and orthogonal PLS-DA identifying major contributor compounds elucidating the differentiation [14].

Beyond characterizing compositional structure, evaluating whether these GSL profiles can consistently distinguish predefined groups provides a complementary measure of biochemical discrimination. ML approaches are well suited to this predictive component because they enable performance to be assessed across complex multivariate data and alternative modelling strategies [15]. In *B. juncea*, multiple algorithms including support vector machine (SVM), random forest (RF), decision tree, generalized linear model, and deep learning approaches have been compared for varietal recognition

using leaf spectral profiles, while boosting algorithms have been applied to predict leaf GSL content, phenolics, flavonoids, anthocyanins, and chlorophyll [16, 17]. In addition, UPLC-ESI-MS/MS profiling coupled with orthogonal PLS-DA and RF identified GSLs and other metabolites associated with varietal and nitrogen-treatment responses in *B. napus* seeds [18]. For metabolite-specific classification, predictive performance alone does not identify the biochemical features underlying class assignment. SHAP provide a model-agnostic framework for attributing predictions to individual features and evaluating both the magnitude and direction of their contributions [19]. Integrating feature attribution with multivariate and predictive analyses therefore provides a coherent approach for evaluating both group discrimination and pinpointing key metabolite contributors.

Accordingly, this study aimed to characterize leaf GSL variation among *B. juncea* accessions representing VM, RM, and SM crop-use types, and to determine the extent to which crop-use differentiation is reflected in their GSL composition. Targeted profiling of 17 GSLs was integrated with univariate and multivariate analyses, pairwise ML classification, and SHAP-based interpretation to evaluate compositional structure, predictive performance, and contribution of individual GSLs to resolution among mustard-types. It was further hypothesized that the three crop-use types would exhibit structured differences in leaf GSL composition despite substantial accession level variation, and that any resulting discrimination would reflect the contributions of multiple GSLs rather than a single dominant compound.

Materials and Methods

2.1 Germplasm acquisition and taxonomic grouping

A total of 42 *Brassica juncea* accessions used in this study were obtained from the National Agrobiodiversity Center (RDA-Genebank) in the Republic of Korea. The collection represented taxonomically diverse materials assigned to *B. juncea* subsp. *integrifolia, B. juncea* subsp. *integrifolia* var. *rugosa, B. juncea* subsp. *juncea, B. juncea* var. *foliosa, B. juncea* var. *integrifolia*, *B. juncea* var. *japonica, B. juncea* var. *longidens, B. juncea* var. *napiformis* and *B. juncea* var. *tumida*. Accessions were classified into three crop-use types according to their taxonomic designation and principal cultivated organ. Following established classifications of *B. juncea* cultivated forms, leaf- and stem-use accessions were consolidated as VM, *napiformis* accessions were classified as RM, and subsp. *juncea* accessions as SM [3]. Accession identity, taxonomical designation, crop-use grouping, and other passport information are provided in Table S1.

2.2 Experimental conditions

Prior to experimental evaluation, the germplasms were cultivated under conditions to maintain accession purity. During greenhouse multiplication in 2022–2023, plants were enclosed with fine mesh material during the reproductive period from February to June to minimize unintended cross-pollination. Seeds obtained from the greenhouse grown plants were subsequently cultivated in the field from

September to November. Plants showing phenotypic heterogeneity within an accession were identified during multiplication and excluded to maintain phenotypic uniformity of the propagated material.

2.3 Sample pre-treatment and glucosinolate extraction

For each accession, three healthy plants were selected as biological replicates. Inner, middle and outer leaves were randomly harvested from each plant, pooled separately, and promptly transferred to polyvinyl bags. The harvested material was freeze for 48 hours using an LP500 vacuum freeze-dryer (Ilshinbiobase Co., Dongducheon, Republic of Korea). After lyophilization, the dried leaves were finely ground into powder form and stored at − 80°C until further processing. Precisely, 0.1 g of freeze-dried leaf powder was mixed with 5 mL methanol (80%) and incubated at 25°C for 30 min, subsequently agitated at 120 rpm for an additional 30 min at same temperature and centrifuged at 14000rpm for 10 min at 4°C. The resulting supernatants were carefully collected into clean vials and retained for subsequent GSL analysis.

2.4 Chromatographic conditions for UPLC-MS/MS analysis

GSL analysis was carried out employing the ACQUITY Ultra-Performance Liquid Chromatography system (Waters, Milford, CT, USA) coupled with Xevo™ TQ-S system (Waters MS Technologies, UK), following the methodology described by [14]. Chromatographic separation was achieved on an ACQUITY UPLC BEH C18 column (1.7 µm, 2.1 × 100 mm) maintained at 35°C. The injection volume was 5 µL, and the mobile phase was delivered at a flow rate of 0.5 mL min^{-1}. Solvent A and solvent B consisted of 0.1% trifluoroacetic acid in water and in methanol, respectively. Elution gradient was programmed according to the solvent ratio (A: B) as follows: 0–7 min; 100:0, 7–10 min, gradually changed to 80:20; 10–11 min, changed to 0:100; 11–15 min, returned to 100:0; and 15–16 min, maintained at 100:0 for column re-equilibration.

2.5 Glucosinolate detection and quantification

Mass spectrometric detection was performed using negative electrospray ionization (ESI) source under multiple reaction monitoring (MRM) conditions. The capillary and cone voltages were set to 3 kV and 54 V, respectively. Individual GSLs were identified by comparing their chromatographic retention times and compound-specific MRM transitions with those of authentic reference standards (Phytolab, Vestenbergsgreuth, Germany; PhytoPlan, Heidelberg, Germany). Quantification was based on external standard calibration using peak area responses and linear regression equations generated for the corresponding GSL standards. Spectrometric data were processed using MassLynx 4.1 software. The MRM transitions, retention times, calibration details for the 17 GSLs retained for downstream analysis are provided in Table S2. GSL concentrations were determined based on these calibration curves and expressed as µmol kg^{-1} sample dry weight (DW).

2.6 Statistical analysis

GSL concentrations were analyzed using linear mixed-effects models with crop-use type as a fixed effect and genotype nested within crop-use type as a random effect. Crop-use type and genotype-within-type

effects were assessed using Satterthwaite-adjusted F-tests and likelihood-ratio tests, respectively, followed by Tukey-adjusted pairwise comparisons at $P < 0.05$. Pearson correlation analysis was used to examine relationships among individual GSLs. HCA was executed on z-score standardized data using Euclidean distance and Ward.D2 linkage. PCA was conducted on centered and unit-variance scaled data, and PLS-DA was performed using the SIMPLS algorithm with two latent components. Variable importance in projection (VIP) scores were calculated to assess the contribution of individual GSL to the PLS-DA model. Statistical and multivariate analyses were carried out in RStudio Version 2024.09.1 + 394 (Posit Software, Boston, MA, USA) using the *lmerTest* [20], *emmeans* [21], *pheatmap* [22], *pls* [23], *ggplot2* [24], and *ggrepel* packages [25].

2.7 Machine learning classification and SHAP analysis

Pairwise ML classification was performed for VM versus RM, VM versus SM, RM versus SM utilizing 17 GSL compounds as predictors. Six algorithms were evaluated, including elastic-net logistic regression (ENLR), linear support vector machine (LSVM), radial support vector machine (RSVM), k-nearest neighbors (kNN), RF, and extreme gradient boosting (XGBoost). Model performance was assessed using 20 repeated stratified three-fold outer cross validation, with two-fold inner cross-validation for hyperparameter tuning, based on accuracy, balanced accuracy, macro-F1 score, Matthews correlation coefficient (MCC), and area under the ROC curve (AUC), where ROC is receiver operating characteristic. SHAP were applied to the selected RM-associated models using model-agnostic Monte Carlo estimation with 1,000 simulations; mean absolute SHAP values were used to rank GSL importance, while signed SHAP values indicated the direction of contribution toward RM classification. All analyses were implemented in Rv4.4.2 using the *caret* [26], *glmnet* [27], *kernlab* [28], *kknn* [29], *randomForest* [30], *xgboost* [31], *pROC* [32], *fastshap* packages [33].

Results

3.1 Variation in glucosinolate composition across *Brassica juncea* germplasm

Leaf GSL profiles varied markedly among the three crop-use types (Table 1). Aliphatic GSLs constituted the predominant fraction of total GSL (TGSL) in all groups, accounting for 98.44% in VM, 98.99% in RM, and 98.07% in SM (Fig. 1A). Aromatic GSLs contributed 1.21, 0.57, and 1.74%, respectively, while indolic GSLs represented 0.34, 0.44, and 0.19%. Sinigrin (SIN) was the major individual GSL across all crop-use types and reached its highest mean concentration in SM (12488.88 μmol kg^{-1} DW) and lowest in RM (7753.15 μmol kg^{-1} DW), with a significant difference between these groups.

Nine of the 17 GSL compounds exhibited significant crop-use type effects. Gluconapin (GNA) and glucobrassicanapin (GBN) are 1202.71 μmol kg^{-1} DW and 19.24 μmol kg^{-1} DW, respectively in VM were significantly higher than in both RM and SM. In addition, SM revealed the highest Glucoiberin (GIB) concentration is 24.87 μmol kg^{-1} DW which was considerably higher than that in RM (9.35 μmol kg^{-1}

DW). Among the ARGSLs, VM and SM accumulated significantly higher gluconasturtiin (GNS) and glucobarbarin (GBB) compared to RM. Distinct configurations were also evident for indolic GSLs; RM contained greater amounts of 4-hydroxyglucobrassicin (4HGBS) and neoglucobrassicin (NGBS) *i.e.* 0.59 and 29.89 µmol kg-1 DW, respectively, whereas 4-methoxyglucobrassicin (4MGBS) was significantly more abundant in VM and SM compared to RM.

Table 1
Leaf glucosinolate concentrations and statistical differences among crop-use types of Indian mustard accessions.

GSL Class	GSL compounds	Crop-use type (Mean ± SE)			ANOVA (*P* value)	
	(umol kg^{-1} DW)	VM	RM	SM	T	G(T)
AGSL	Sinigrin	10666.18 ± 498.48^{ab}	7753.15 ± 603.86^{b}	12488.88 ± 473.09^{a}	*	***
	Gluconapin	1202.71 ± 481.23^{a}	65.90 ± 8.28^{b}	84.58 ± 15.31^{b}	**	***
	Glucobrassicanapin	19.24 ± 8.38^{a}	2.16 ± 0.67^{ab}	0.89 ± 0.58^{b}	*	***
	Epiprogoitrin	5.03 ± 1.96^{a}	3.02 ± 1.75^{a}	1.46 ± 0.64^{a}	ns	***
	Progoitrin	4.74 ± 1.80^{a}	3.80 ± 2.03^{a}	1.47 ± 0.60^{a}	ns	***
	Glucoraphasatin	0.16 ± 0.04^{a}	0.02 ± 0.00^{a}	0.02 ± 0.01^{a}	ns	**
	Glucoiberin	15.89 ± 2.02^{ab}	9.35 ± 2.12^{b}	24.87 ± 3.46^{a}	*	***
	Glucoraphenin	0.18 ± 0.10^{a}	ND	0.01 ± 0.00^{a}	ns	***
	Glucoraphanin	2.59 ± 1.39^{a}	0.10 ± 0.01^{a}	0.09 ± 0.02^{a}	ns	***
	Glucocheirolin	776.85 ± 44.54^{a}	502.52 ± 46.99^{a}	1051.98 ± 41.61^{a}	ns	ns
ARGSL	Glucotropaeolin	1.44 ± 0.12^{a}	1.45 ± 0.29^{a}	1.42 ± 0.14^{a}	ns	***
	Gluconasturtiin	151.82 ± 22.04^{a}	44.80 ± 7.57^{b}	236.71 ± 60.17^{a}	**	***
	Glucobarbarin	3.16 ± 0.33^{a}	1.42 ± 0.16^{b}	4.47 ± 0.93^{a}	**	***
	Sinalbin	0.01 ± 0.00^{a}	0.01 ± 0.00^{a}	0.02 ± 0.01^{a}	ns	ns
INGSL	4-Hydroxyglucobrassicin	0.16 ± 0.03^{a}	0.59 ± 0.15^{b}	0.08 ± 0.04^{a}	**	***
	4-Methoxyglucobrassicin	33.13 ± 4.23^{a}	7.01 ± 1.42^{b}	16.56 ± 2.42^{a}	***	***
	Neoglucobrassicin	10.92 ± 1.12^{a}	29.89 ± 7.91^{b}	10.09 ± 1.91^{ab}	*	***

GSL Class	GSL compounds	Crop-use type (Mean ± SE)			ANOVA (*P* value)	
	(umol kg^{-1} DW)	VM	RM	SM	T	G(T)
TGSL	Total Glucosinolates	12894.20 ± 313.55[a]	8425.19 ± 667.51[b]	13923.61 ± 543.86[a]	***	***

SE, standard error; VM, vegetative mustard; RM, root mustard; SM, seed mustard, ALGSL, aliphatic glucosinolate; ARGSL, aromatic glucosinolate; INGSL, indolic glucosinolate; TGSL, total glucosinolate. Different alphabets indicate significant differences based on Tukey-adjusted comparisons (P < 0.05). ns, non-significant; *, **, and ***represents P < 0.05, 0.01, and 0.001, respectively; ND, not detected.

TGSL concentration also varied significantly among crop-use types (Table 1; Fig. 1B). VM and SM belonged to the same statistical group, with mean TGSL value of 12894.20 ± 313.55 and 13923.61 ± 543.86 μmol kg^{-1} DW, respectively, while RM showed considerably lower concentration of 8425.19 ± 667.51 μmol kg^{-1} DW. Significant genotype-within-type effect occurred for 15 of the 17 individual GSLs, demonstrating pronounced accession level variation within each mustard types.

3.2 Hierarchical clustering and principal component analysis of glucosinolate profiles

HCA partitioned the accessions based on GSL composition (Fig. 2A), highlighting accession-level diversity across the germplasm panel. Notably, IT297448 and IT297449 from India, and IT297458 from Thailand formed a distinct cluster characterized by comparatively high relative abundance of GNA, GBN, epiprogoitrin (EPI), progoitrin (PRO), glucoraphasatin (GRH), glucoraphenin (GRE), and glucoraphanin (GRA), together with lower SIN and glucocheirolin (GCH). The remaining accessions were distributed among broader clusters exhibiting diverse combinations of aliphatic, aromatic, and indolic GSLs. Pearson correlation analysis further supported coordinated variation among several of these compounds, including positive associations between EPI, GRE, SIN, GNA, and GNS with PRO, GRA, GCH, GBN, and GBB, respectively (Fig S1).

PCA further resolved the major axes of variation in GSL composition among the 42 accessions (Fig. 2B). PC1 and PC2 explained 43.3% and 15.6% of the total variance, respectively, accounting for 58.9% cumulatively. RM accessions occupied a comparatively distinct region, predominantly at negative PC2 scores, while SM accessions are concentrated toward positive PC2 scores and overlapped with the upper portion of the broader VM distribution. VM showed the greatest within-group dispersion. Importantly, the three VM accessions IT297448, IT297449, and IT297458 (*Brassica juncea* var. *japonica*) were markedly displaced toward positive PC1, consistent with their distinct GSL profiles observed by HCA.

The loading structure provided a compositional basis for these accession distributions (Fig. 2C). GRA, GRE, GRH, GBN, and PRO contributed predominantly toward positive PC1, whereas SIN, GCH, NGBS, and glucotropaeolin (GTL) were oriented toward negative PC1. Along PC2, GBB, GNS, 4MGBS, and GIB showed positive loadings, with NGBS and 4HGBS contributing in the opposite direction. This loading pattern corresponded with the crop-use distribution, placing RM closer to NGBS and 4HGBS and SM closer to GBB, GNS, 4MGBS, and GIB. A complementary PLS-DA displayed a broadly comparable crop-use associated distribution, with corresponding score and VIP plots provided in Fig S2.

3.3 Comparative machine learning classification of Indian mustard crop-use types

Pairwise ML analysis revealed pronounced differences in the predictability of the Indian mustard types. Across six evaluated algorithms, VM versus RM and RM versus SM showed consistently stronger classification performance (Fig. 3). LSVM provided the strongest performance for VM versus RM, with accuracy, balanced accuracy, and macro-F1 of 0.99, together with MCC and AUC of 0.98 and 1.00, respectively. Similarly, RF showed high performance for RM versus SM, while its performance for VM versus SM was substantially lower, particularly for balanced accuracy and MCC.

The normalized confusion matrices further demonstrated the class-specific performance of the selected models (Fig. 3B-D). LSVM almost completely resolved VM from RM, correctly assigning all VM observations and 97% of RM observations. For RM and SM, RF showed strong discrimination with complete recovery of RM and 99% true prediction of SM. In contrast, VM and SM remained difficult to distinguish. Although 95% of VM observations were correctly identified, only 28% of SM observations were recovered. This patterns were also reflected in the ROC analysis, where VM versus RM and RM versus SM each attained an AUC of 1.00 except for VM versus SM (Fig. 3E-G).

3.4 SHAP analysis of glucosinolate contributions to root mustard type classification

Feature attribution differed between the two RM-associated classification models, while retaining a substantial set of common contributors (Fig. 4). In the VM versus RM comparison LSVM model, 4HGBS showed the highest mean absolute SHAP value, followed by SIN, 4MGBS, GCH, and GNA with NGBS, GNS, GBB, GRH, and Sinalbin (SNB) contributing to a lesser extent (Fig. 4A). The corresponding SHAP distribution showed distinct directional effects among these compounds (Fig. 4B). Higher 4HGBS and NGBS levels, together with lower levels of several prominent ALGSLs and ARGSLs generally shifted predictions toward RM. This contrast therefore reflected a mixed attribution pattern, in which RM classification was characterized by both enrichments of selected indolic features and relative depletion of several otherwise influential GSLs.

A partly overlapping but differently ranked feature profile was obtained for the RM versus SM RF model (Fig. 4C, D). The feature importance was concentrated most strongly in GCH, GBB, and SIN with mean absolute SHAP values of approximately 0.06. Similar to VM versus RM, directional SHAP patterns

exhibited positive RM-directed predictions were associated with higher values and several INGSLs and lower values of ALGSLs and ARGSLs relative of SM. Cross contrast comparison further revealed that eight out of ten highest ranked GSLs were common to both models, indicating substantial stability in the principal predictive features despite differences in their relative importance (Table S3). This contrast specific redistribution was evident for GNA, which contributed more strongly to VM versus RM classification, and GIB, which was more prominent in discriminating RM from SM.

Discussion

Brassica juncea comprises agronomically distinct forms that have undergone extensive diversification during domestication and selection. Previous population-genomic studies have shown that the species diversified along multiple evolutionary trajectories, leading to the differentiation of genetically distinct cultivated forms including root, seed, leafy and stem mustard subspecies [3]. In the present study, leaf GSL profiles were compared across 42 accessions representing three different crop-use types. This framework enabled assessment of whether agronomic diversification in *B. juncea* is reflected in variation in leaf GSL composition.

4.1 Leaf glucosinolate patterns associated with crop-use diversification in *B. juncea*.

The dominance of ALGSLs, particularly SIN, across all three mustard groups indicates the crop-use differentiation has occurred within a broadly conserved leaf GSL background. Similar biochemical patterns have been reported in a large *B. juncea* leaf germplasm collections, with SIN identified as the predominant GSL in 210 Korean leaf-mustard accessions and in broad panel of 355 leaf-mustard accessions [10, 11]. Both studies also documented pronounced accession level variation in individual and total GSL concentrations, highlighting the considerable biochemical diversity retained with *B. juncea* germplasm. Consistent with this pattern, significant genotype-within-type effects were detected for most GSL in this study (Table 1), indicating marked accession level heterogeneity within each crop-use group rather than chemically uniform group-specific GSL profiles.

Against the shared SIN-dominated background, RM displayed the most distinct compositional profile, characterized by lower TGSL (Fig. 1B) and reduced accumulation of several prominent ALGSLs and ARGSLs, but comparatively higher 4HGBS and NGBS (Table 1). Comparable variation in INGSLs, including 4HGBS, 4MGBS, and NGBS, has been documented in *B. juncea* leaf accessions [7]. Together with the contrasting GNA and GBN levels in VM and GIB, GNS, and GBB in SM, these patterns demonstrate that differentiation among mustard types involves changes in the relative balance of individual GSLs rather than merely differences in total GSL accumulation. This interpretation is particularly relevant because GSL abundance and composition vary strongly among *B. juncea* tissues and developmental stages [34, 35]. In addition, as the present comparison was based exclusively on leaf tissue, the observed contrasts reflect variation in leaf GSL composition among the three mustard types rather than differences attributable to their respective harvested organs.

4.2 Multivariate clustering of accessions across different mustard types

The multivariate analyses revealed that variation in leaf GSL composition was not organized into three completely discrete crop-use groups. Instead, HCA and PCA showed partial structuring, with RM accessions occupying a comparatively distinct position, while SM remained extensively embedded within the broader VM distribution (Fig. 2). Similar accession-dependent differentiation in leaf GSL content has been mentioned in *Brassica* germplasms, where multivariate analyses resolved chemically divergent accessions and groups based on contrasting GSL accumulation [36, 37]. However, our results suggest that GSL-based clustering differed among mustard types, with robust compositional separation of RM than between VM and SM.

The comparatively distinct placement of RM is consistent with recently reported evidence specifying root type mustard (*napiformis*) as a differentiated evolutionary lineage within *B. juncea* [38]. Although genomic rather than metabolic, the differentiation provides evolutionary context for stronger GSL separation observed in RM. Because RM in the present panel corresponds to var. *napiformis* and SM to subspecies *juncea*, crop-use and taxonomic identity are partly confounded; the RM separation should therefore be interpreted as root mustard/*napiformis*-associated feature rather than an independent crop-use effect.

Notably, three VM accessions, IT297448, IT297449, and IT297458 (*B. juncea* var. *japonica*) formed a distinct cluster in HCA and positioned in the positive extreme of PC1 in PCA, suggesting that accession level taxonomic background may also contribute to the multivariate organization of leaf GSL profiles. Recent metabolomics characterization of genetically diverse *B. rapa* and *B. oleracea* accessions revealed pronounced inter- and intraspecific chemical diversity and demonstrated that metabolic differentiation can persist both among taxonomical groups and among accessions within the same crop lineage [39]. This supports the view that the distinct positioning of the three *B. juncea* var. *japonica* accessions may reflect accession-specific biochemical differences superimposed on the broader crop-use structure.

4.3 Predictive discrimination of mustard types and key glucosinolate signatures

The pairwise ML analysis quantified the asymmetric pattern observed by PCA. VM versus RM and RM versus SM were discriminated with consistently high performance, however VM versus SM showed substantially weaker classification, including lower MCC and frequent assignment of SM accessions to VM (Fig. 3). This contrast-dependent predictability reinforces the multivariate evidence that RM carries a more distinctive leaf GSL composition, while the GSL boundaries between VM and SM remain comparatively diffused. Previous studies have similarly demonstrated that metabolite composition can support ML-based discrimination among biological or agronomic groups. In *Pimpinella anisum*, physicochemical and metabolic variables contributed to RF classification among accessions and

experimental groups [40], while metabolite-driven ML successfully recognized pigmented onion types and identified influential compounds underlying their classification [41].

SHAP attribution further indicated that the strong RM-related classifications were not dependent on a single GSL but to a recurrent combination of compounds spanning different GSL classes (Fig. 4). Eight GSLs comprising SIN, GCH, 4HGBS, 4MGBS, GBB, GNS, NGBS, and SNB consistently ranked among the most influential variables in both RM-associated comparisons. In contrast, GNA and GRH showed greater importance in VM versus RM, besides GIB and GBN were more prominent in RM versus SM. The prominence of 4HGBS and NGBS is consistent with their comparatively higher accumulation in RM, while the contribution of several ALGSLs and ARGSLs reflects the broader compositional shifts identified in the univariate analyses. Thus, the convergence suggests that RM discrimination arises from a coordinated shift in GSL signatures rather than from the presence or absence of an isolated metabolite. SHAP has similarly been used in metabolomics-based ML to identify phytochemical features contributing to species resolution in *Zanthoxylum* [42]. However, because attribution among correlated predictors can be redistributed across related variables, differences in SHAP rank should not be interpreted as independent biological effects of individual GSLs [43]. These recurrently influential GSL signatures are therefore more appropriately considered as candidate GSL discriminators of RM within the present germplasm panel rather than validated biomarkers.

4.4 Limitations and future perspectives

Despite the promising discrimination of mustard types based on leaf GSL profiles in this study, several aspects warrant further investigation. Broader accession coverage would provide a stronger assessment of how consistently the identified patterns extend across the wider diversity of *B. juncea*. Environmental stability also requires consideration, as GSL accumulation in Indian mustard is influenced by genotype, environment, and their interaction [44], and nutrient availability and stress can alter leaf GSL composition [45]. Accordingly, multi-year and multi-location evaluation would be useful for assessing the stability of these signatures. In addition, although repeated internal cross-validation provided robust estimates of predictive performance, validation in independent germplasm would further establish the transferability of the recurrent GSL discriminators. The present targeted profiling focused on intact GSLs and did not include downstream hydrolysis products, which may provide complementary biochemical information because their formation can vary among *B. juncea* cultivars [46]. To address these limitations, future studies incorporating these approaches would help determine whether the GSL signatures discerning mustard types are reproducible across independent germplasm and environments and whether they are accompanied by consistent differences in downstream GSL metabolism.

Conclusions

This study revealed that leaf GSL profiling can provide complementary biochemical basis for distinguishing major crop-use forms and taxonomical relevance of *B. juncea*. RM accessions exhibited the most distinct GSL composition and showed consistent separation from VM and SM across multivariate and ML analyses. VM and SM displayed substantially greater overlap in their leaf GSL

profiles. In particular, recurrent contributions of 4HGBS, NGBS, SIN, GCH and related compounds further defined a coordinated GSL signature associated with RM differentiation. These compounds represent candidate GSL discriminators that require validation across broader germplasm and environments. Collectively, the targeted GSL-based framework may facilitate germplasm evaluation and characterization, including the prioritization of contrasting genetic resources aiming subsequent breeding strategies and mustard improvement programs.

Abbreviations

GSL
Glucosinolate
TGSL
Total glucosinolate
VM
Vegetative mustard
RM
Root mustard
SM
Seed mustard
PCA
Principal component analysis
PLS-DA
Partial least squares discriminant analysis
HCA
Hierarchical clustering analysis
UPLC-MS/MS
Ultra-performance liquid chromatography -tandem mass spectrometry
ESI
Electrospray ionization
MRM
Multiple reaction monitoring
VIP
Variable importance in projection
ENLR
Elastic-net logistic regression
LSVM
Linear support vector machine
RSVM
Radial support vector machine
kNN

k-nearest neighbors

RF

Random forest

XGBoost

Extreme gradient boosting

MCC

Matthews correlation coefficient

AUC

Area under the receiver operating characteristic curve

SHAP

Shapley additive explanations

SIN

Sinigrin

GNA

Gluconapin

GBN

Glucobrassicanapin

EPI

Epiprogoitrin

PRO

Progoitrin

GIB

Glucoiberin

GCH

Glucocheirolin

GRA

Glucoraphanin

GRE

Glucoraphenin

GRH

Glucoraphasatin

SNB

Sinalbin

GTL

Glucotropaeolin

GNS

Gluconasturtiin

NGBS

Neoglucobrassicin

4HGBS

4-hydroxyglucobrassicin
4MGBS
4-methoxyglucobrassicin

Declarations

Data availability

The dataset analyzed used and/or analyzed in this study will be made available upon reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Data availability

The dataset analyzed used and/or analyzed in this study will be made available upon reasonable request.

Competing interests

The authors declare no competing interests.

Funding

This research was funded by the Research Program for Agricultural Science and Technology Development (Project No. PJ01425501/RS-2019-RD007776), National Institute of Agricultural Sciences, Rural Development Administration (RDA), Republic of Korea.

Acknowledgements

The authors would like to acknowledge researcher Kingsley Ochar for his assistance with this research study during his tenure at Rural Development Administration.

Author information

Authors and Affiliations

National Institute of Agrobiodiversity Center, National Institute of Agricultural Sciences, Rural Development Administration, Jeonju 5487, Republic of Korea

Sreeparna Chowdhury, Kanivalan Iwar & Seong-Hoon Kim

Contributions

S.C.: Writing – Original draft, Visualization, Methodology, Formal analysis, Data curation, Validation, Analysis, Conceptualization. **K.I.:** Investigation, Data curation, Analysis. **S.-H.K:** Writing-review & editing, Supervision, Resources, Funding acquisition, Project administration, Data curation.

Corresponding author

Correspondence to Seong-Hoon Kim.

References

1. Sun B, Tian Y-X, Jiang M, Yuan Q, Chen Q, Zhang Y, Luo Y, Zhang F, Tang H-R. Variation in the main health-promoting compounds and antioxidant activity of whole and individual edible parts of baby mustard (*Brassica juncea* var. *gemmifera*). RSC Adv. 2018;8(59):33845–54. https://doi.org/10.1039/c8ra05504a.
2. Parveen M, Khan AU, Shahbaz HM, Raza SQ, Khaliq M. Nutritional composition, phytochemical profile, and in-vitro antioxidant and antimicrobial activities of mustard greens (*Brassica juncea*) leaves. Int J Food Prop. 2026;29(1):2713292. https://doi.org/10.1080/10942912.2026.2713292.
3. Kang L, Qian L, Zheng M, Chen L, Chen H, Yang L, You L, Yang B, Yan M, Gu Y, et al. Genomic insights into the origin, domestication and diversification of *Brassica juncea*. Nat Genet. 2021;53(9):1392–402. https://doi.org/10.1038/s41588-021-00922-y.
4. Zhang D, Yu H, Gao L, Wang J, Dong H, Guo Y, Hu S. Genetic Diversity in Oilseed and Vegetable Mustard (*Brassica juncea* L.) Accessions Revealed by Nuclear and Mitochondrial Molecular Markers. Agronomy. 2023;13(3):919. https://doi.org/10.3390/agronomy13030919.
5. Cools K, Terry LA. The effect of processing on the glucosinolate profile in mustard seed. Food Chem. 2018;252. https://doi.org/10.1016/j.foodchem.2018.01.096. :343 – 48.
6. Nambiar DM, Kumari J, Augustine R, Kumar P, Bajpai PK, Bisht NC. GTR1 and GTR2 transporters differentially regulate tissue-specific glucosinolate contents and defence responses in the oilseed crop *Brassica juncea*. Plant Cell Environ. 2021;44(8):2729–43. https://doi.org/10.1111/pce.14072.
7. Huang H, Wang J, Mao S, Wu Q, Tian Y, Wang F, Wang P, Huang K, Wu Q. Variation Characteristics of Glucosinolate Contents in Leaf Mustard (*Brassica juncea*). Agronomy. 2022;12(10):2287. https://doi.org/10.3390/agronomy12102287.
8. Zhang C, Di H, Lin P, Wang Y, Li Z, Lai Y, Li H, Sun B, Zhang F. Genotypic variation of glucosinolates and their breakdown products in mustard (*Brassica juncea*) seeds. Sci Hortic. 2022;294:110765. https://doi.org/10.1016/j.scienta.2021.110765.
9. Li Z, Di H, Cheng W, Zhang Y, Ren G, Ma J, Yang J, Huang Z, Tang Y, Zheng Y, et al. Variation in health-promoting compounds and antioxidant activities in mustard (*Brassica juncea*) sprouts. Sci Hortic.

2023;309:111673. https://doi.org/10.1016/j.scienta.2022.111673.

10. Kim HW, Ko HC, Baek HJ, Cho SM, Jang HH, Lee YM, Kim JB. Identification and quantification of glucosinolates in Korean leaf mustard germplasm (*Brassica juncea* var. *integrifolia*) by liquid chromatography–electrospray ionization/tandem mass spectrometry. Eur Food Res Technol. 2016;242(9):1479–84. https://doi.org/10.1007/s00217-016-2648-6.
11. Assefa AD, Kim S-H, Ko HC, Ro N, Subramanian P, Chung Y-J, Lee Y-H, Hahn B-S, Rhee J-H. Leaf Mustard (*Brassica juncea*) Germplasm Resources Showed Diverse Characteristics in Agro-Morphological Traits and Glucosinolate Levels. Foods. 2023;12(23):4374.
12. Worley B, Powers R. Multivariate Analysis in Metabolomics. Curr Metabolomics. 2013;1(1):92–107. https://doi.org/10.2174/2213235x11301010092.
13. Liang X, Lee HW, Li Z, Lu Y, Zou L, Ong CN. Simultaneous quantification of 22 glucosinolates in 12 Brassicaceae vegetables by hydrophilic interaction chromatography–tandem mass spectrometry. ACS omega. 2018;3(11):15546–53.
14. Kim S-H, Lee G-A, Subramanian P, Hahn B-S. Quantification and diversity analyses of major glucosinolates in conserved chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*) germplasms. Foods. 2023;12(6):1243.
15. Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. Trends Plant Sci. 2014;19(12):798–808. https://doi.org/10.1016/j.tplants.2014.08.004.
16. Sohn SI, Pandian S, Oh YJ, Zinia Zaukuu JL, Lee YH, Shin EK. Discrimination of *Brassica juncea* Varieties Using Visible Near-Infrared (Vis-NIR) Spectroscopy and Chemometrics Methods. Int J Mol Sci. 2022;23(21). https://doi.org/10.3390/ijms232112809.
17. Yoon HI, Lee H, Yang J-S, Choi J-H, Jung D-H, Park YJ, Park J-E, Kim SM, Park SH. Predicting Models for Plant Metabolites Based on PLSR, AdaBoost, XGBoost, and LightGBM Algorithms Using Hyperspectral Imaging of *Brassica juncea*. Agriculture. 2023;13(8):1477. https://doi.org/10.3390/agriculture13081477.
18. Wang C, Li Z, Zhang L, Gao Y, Cai X, Wu W. Identifying Key Metabolites Associated with Glucosinolate Biosynthesis in Response to Nitrogen Management Strategies in Two Rapeseed (*Brassica napus*) Varieties. J Agric Food Chem. 2022;70(2):634–45. https://doi.org/10.1021/acs.jafc.1c06472.
19. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. https://doi.org/10.48550/arXiv.1705.07874.
20. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Softw. 2017;82:1–26.
21. Lenth R, emmeans. Estimated Marginal Means, aka Least-Squares Means_. R package version. 2023;18:5.
22. Kolde R. Pretty Heatmaps [R package pheatmap version 1.0. 13]. CRAN: Contributed Packages. 2025;10.

23. Mevik B-H, Wehrens R. The pls package: principal component and partial least squares regression in R. J Stat Softw. 2007;18:1–23.
24. Wickham H. Data analysis. ggplot2: elegant graphics for data analysis. Springer; 2016. pp. 189–201.
25. Slowikowski K, Schep A, Hughes S, Lukauskas S, Irisson J-O, Kamvar ZN, Ryan T, Christophe D, Hiroaki Y, Gramme P. Package ggrepel. Automatically position non-overlapping text labels with 'ggplot2. 2018:1–9.
26. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008;28(5):1–26. https://doi.org/10.18637/jss.v028.i05.
27. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22. https://doi.org/10.18637/jss.v033.i01.
28. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab - An S4 Package for Kernel Methods in R. J Stat Softw. 2004;11(9):1–20. https://doi.org/10.18637/jss.v011.i09.
29. Hechenbichler K, Schliep K. Weighted k-nearest-neighbor techniques and ordinal classification. 2004.
30. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.
31. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining: 2016. 785 – 94.
32. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77. https://doi.org/10.1186/1471-2105-12-77.
33. Greenwell B. fastshap: fast approximate Shapley values. CRAN: Contributed Packages; 2019.
34. Bellostas N, Sørensen JC, Sørensen H. Profiling glucosinolates in vegetative and reproductive tissues of four *Brassica* species of the U-triangle for their biofumigation potential. J Sci Food Agric. 2007;87(8):1586–94. https://doi.org/10.1002/jsfa.2896.
35. Púčiková V, Rohn S, Hanschen FS. Glucosinolate Accumulation and Hydrolysis in Leafy *Brassica* Vegetables Are Influenced by Leaf Age. J Agric Food Chem. 2023;71(30):11466–75. https://doi.org/10.1021/acs.jafc.3c01997.
36. Rhee J-H, Choi S, Lee J-E, Hur O-S, Ro N-Y, Hwang A-J, Ko H-C, Chung Y-J, Noh J-J, Assefa AD. Glucosinolate content in *Brassica* genetic resources and their distribution pattern within and between inner, middle, and outer leaves. Plants. 2020;9(11):1421.
37. Wang J, Yu S, Ren X, Wang Y, Wang Z, Meng Q, Huang Y, Wang Y. Integrative Analyses of Metabolites and Transcriptome Reveal the Metabolic Pattern of Glucosinolates in Potherb Mustard (*Brassica juncea* var. *multiceps*). Plants. 2024;13(17):2481. https://doi.org/10.3390/plants13172481.
38. Wang H, Cai X, Guan Z, Wu J, Peng L, Li W, Rao L, Yang S, Zhang Z, Zhang X, et al. Population genomic insights into the domestication of Brassica juncea var. tumida. Hortic Res. 2026;13(2). https://doi.org/10.1093/hr/uhaf298.

39. Le Boulch P, Missinou AA, Boutet S, Perreau F, Manzanares-Dauleux MJ, Falentin C, Filangi O, Lariagon C, Bouchereau A, Rousseau-Gueutin M, et al. Metabolomic analysis of a core collection of *Brassica rapa* and *Brassica oleracea* unveils unexpected chemical diversity with potential applications in chemical ecology and breeding. BMC Plant Biol. 2026;26(1):569. https://doi.org/10.1186/s12870-026-08269-4.
40. de Oliveira BR, Zuffo AM, dos Santos Silva FC, Steiner F, AlGarawi AM, Okla MK, Nhs M, Alhaj Hamoud Y, Josko I, Sheteiwy MS, et al. Random forest algorithms: a tool to identify the impact of arbuscular mycorrhizal fungi inoculation, seed maturation stage and geographic diversity of *Pimpinella anisum* L. accessions on the physicochemical composition of seeds. BMC Plant Biol. 2025;25(1):608. 10.1186/s12870-025-06536-4.
41. Cheng K, Xiao J, He J, Yang R, Pei J, Jin W, Abd El-Aty AM. Unraveling volatile metabolites in pigmented onion (*Allium cepa* L.) bulbs through HS-SPME/GC–MS-based metabolomics and machine learning. Front Nutr. 2025;12–2025. https://doi.org/10.3389/fnut.2025.1582576.
42. Yang S, Li Z, Li W, Li C, Yang X, Zhao Y, Liu R. Unraveling the variations and relationships between phytochemical constituents and anti-inflammatory potentials of three *Zanthoxylum* species through metabolomics and explainable machine learning. Ind Crops Prod. 2024;212:118281. https://doi.org/10.1016/j.indcrop.2024.118281.
43. Aas K, Martin J, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell. 2021;298:103502. https://doi.org/10.1016/j.artint.2021.103502.
44. Chauhan JS, Singh KH, Meena SS, Singh M. Environmental effects on oil, fatty acid profile, protein and glucosinolate content in Indian mustard (*Brassica juncea*). Indian J Agricultural Sci. 2011;81(7). 10.56093/ijas.v81i7.7383.
45. Almuziny M, Decker C, Wang D, Gerard P, Tharayil N. Nutrient Supply and Simulated Herbivory Differentially Alter the Metabolite Pools and the Efficacy of the Glucosinolate-Based Defense System in *Brassica* Species. J Chem Ecol. 2017;43(2):129–42. https://doi.org/10.1007/s10886-016-0811-y.
46. Ko D-Y, Seo S-M, Lee Y-H, Gil CS, Lee H, Ku K-M. Turning glucosinolate into allelopathic fate: investigating allyl isothiocyanate variability and nitrile formation in eco-friendly *Brassica juncea* from South Korea. Sci Rep. 2024;14(1):15423. https://doi.org/10.1038/s41598-024-65938-w.

Figures

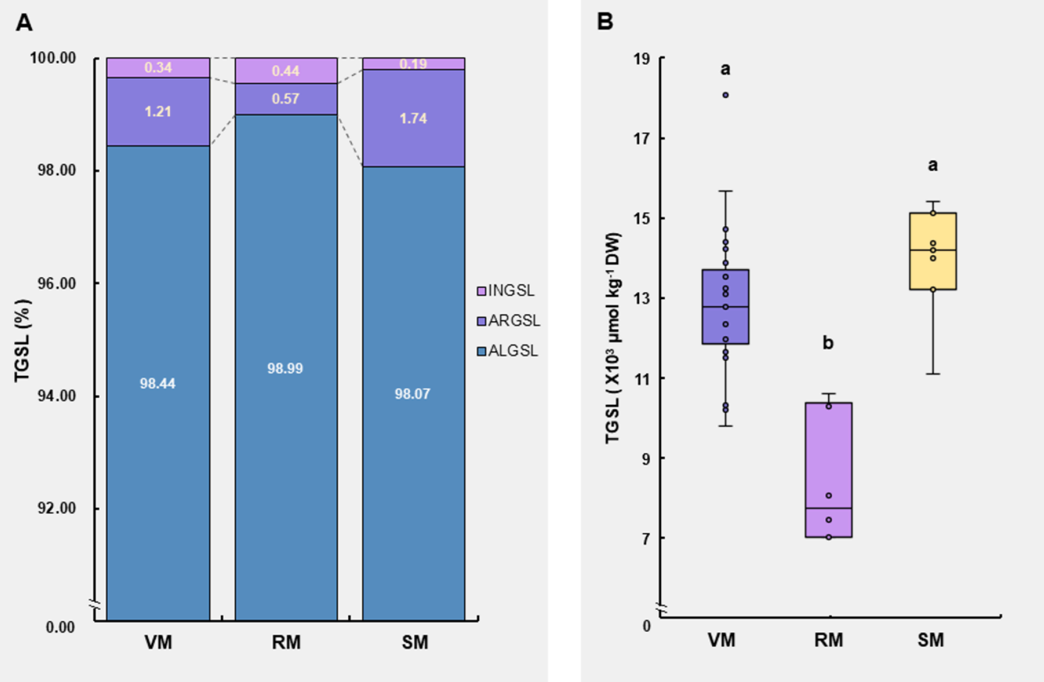


Figure 1

Relative glucosinolate class composition and total glucosinolate concentrations among mustard types. (A) Stacked bar plots representing the relative contributions of aliphatic (ALGSL), aromatic (ARGSL), and indolic (INGSL) glucosinolates to total glucosinolate (TGSL) in vegetative (VM), root (RM), and seed (SM) mustard types, respectively. (B) Distribution of TGSL concentrations among individual accessions within each crop-use type. Different lowercase letters indicate significant differences based on Tukey's-adjusted pairwise comparison ($P < 0.05$).

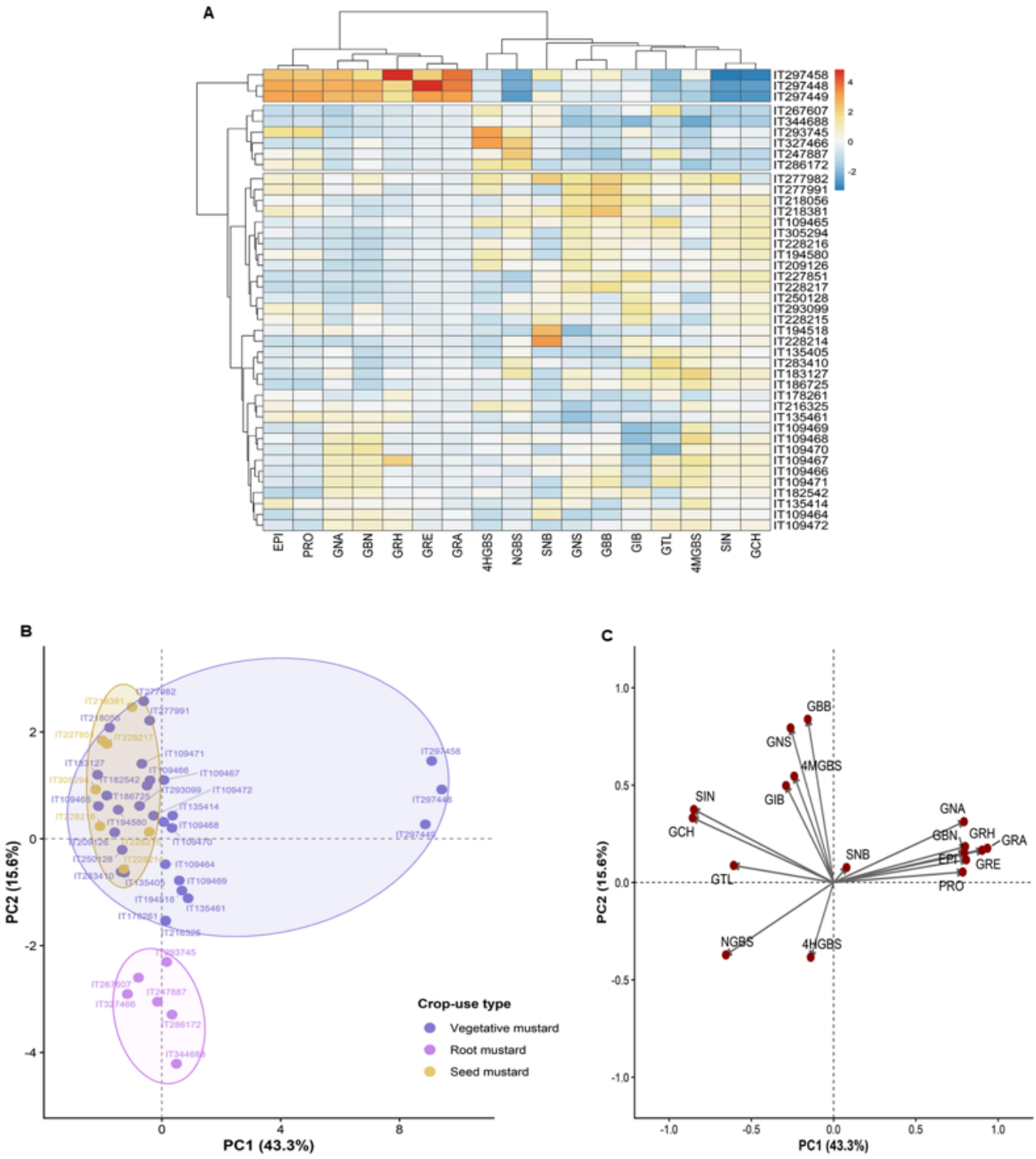


Figure 2

HCA-based clustering, PCA score plot, and loading plot of glucosinolates. (A) Hierarchical heatmap of 17 individual GSLs of the 42 Indian mustard accessions. (B) PCA score plot showing distribution of vegetative (VM), root (RM), and seed (SM) mustard accessions. (C) Loading plot depicting the contribution and the direction of individual GSLs associated with accession distribution along PC1 and PC2.

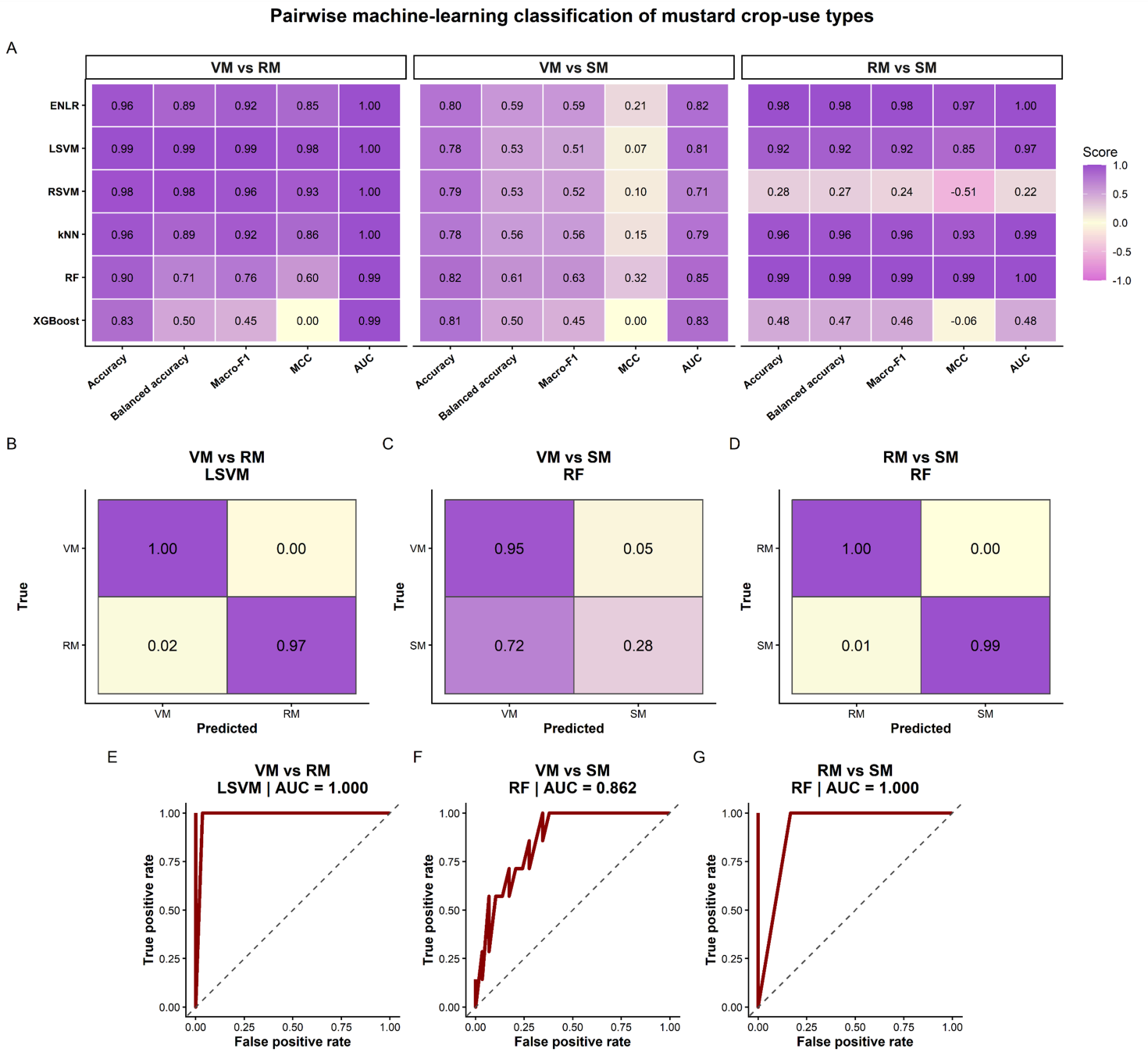


Figure 3

Comparative performance of pairwise ML classification in *B. juncea*. (A) Performance of six ML models (ENLR, LSVM, RSVM, kNN, RF, and XGBoost) for VM versus RM, VM versus SM, and RM versus SM classification. Heatmap values represent mean accuracy, balanced accuracy, macro-F1, MCC, and AUC across repeated stratified cross-validation. (B-D) Row-normalized confusion matrices for the best performing model in each comparison, with rows and columns indicating true classes and predicted classes, respectively. (E-G) ROC curves for the corresponding selected models, with AUC values shown above each panel. VM, vegetative mustard; RM, root mustard; SM, seed mustard.

Supplementary Files

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

- Table1.xlsx
- SupplementaryBJSC.docx