High-Throughput Seed Phenotyping and GWAS Uncover Key Genetic Variants Influencing Seed Quality in Leymus chinensis

Shuangyan Chen

sychen@ibcas.ac.cn

Institute of Botany, Chinese Academy of Sciences

Lei Tian

Chuifan Kong

Shuyi Hu

Xiaoyun Han

Qun Sun

Liqin Cheng

Institute of Botany, Chinese Academy of Sciences

Gongshe Liu

Article

Keywords: Leymus chinensis, High-hroughput phenotyping, GWAS, Seed traits, Molecular breeding

Posted Date: December 5th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7733701/v1

Additional Declarations: There is **NO** Competing Interest.

High-Throughput Seed Phenotyping and GWAS Uncover Key Genetic Variants Influencing Seed Quality in *Leymus chinensis*

Lei Tian[1,2,3], Chuifan Kong[1,2,3], Shuyi Hu[1,2,3], Xiaoyun Han[1,2,3], Qun Sun[4], Liqin Cheng[2], Gongshe Liu[2*], Shuangyan Chen[1,2*]

[1]State Key Laboratory of Forage Breeding-by-Design and Utilization, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[2]Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[3]University of Chinese Academy of Sciences, Beijing 100049, China

[4]College of Agronomy and Biotechnology, China Agricultural University/The Innovation Center (Beijing) of Crop Seeds Whole-Process Technology Research of Ministry of Agriculture and Rural Affairs, Beijing 100193, China

*Correspondence: Shuangyan Chen (sychen@ibcas.ac.cn)，Gongshe Liu (liugs@ibcas.ac.cn)

ABSTRACT

Leymus chinensis (Trin.) Tzvel. (sheepgrass) is an important forage species, yet the relationships between seed phenotypic traits, agronomic performance, and their underlying genetic mechanisms remain unclear. In this study, we utilized the AIseed high-throughput phenotyping platform to systematically analyze 54 image-based traits (i-traits)—encompassing morphology, color, and texture—in 262 dehusked seeds of sheepgrass. Coupled with 50K single nucleotide polymorphism (SNP) chip genotyping data, we performed a genome-wide association study (GWAS) to elucidate genetic correlations among seed phenotypic traits. Elastic net regression was employed to identify informative phenotypic predictors, revealing significant associations between seed size, seed coat texture, and color with hundred-seed weight (HGW), hundred-seed weight without glumes (HGWwg), and germination rate (GR). Additionally, a germplasm screening approach based on principal component analysis (PCA) achieved a 71% accuracy rate in predicting high-germination germplasm and identified 10 germplasm lines with superior comprehensive performance. GWAS identified several SNPs significantly associated with seed color and morphology, mainly on chromosomes Lc2Xm and Lc6Xm. KEGG analysis highlighted the roles of phenylpropanoid and flavonoid biosynthesis pathways, with candidate genes such as *PAL*, *PER18*, *PER50*, *BGLU16*, *BACOVA_02659*, and *ANR* implicated. This study offers effective phenotypic screening strategies and valuable genetic resources for the molecular breeding of sheepgrass.

KEYWORDS: *Leymus chinensis*, High-hroughput phenotyping, GWAS, Seed traits, Molecular breeding

INTRODUCTION

Seed quality traits are fundamental determinants of plant reproductive success and agricultural productivity. Extensive research has demonstrated that seed morphological characteristics—such as size, color, and surface texture—play critical roles in influencing dispersal efficiency (Liu et al., 2013), regulating dormancy (Krzyszton et al., 2024), and enhancing germination performance (Mahajan et al., 2018), thereby shaping crop yield potential (Hu et al., 2023; Oh et al., 2024). *Leymus chinensis* (Trin.) Tzvel., commonly known as sheepgrass, is a perennial grass species of considerable ecological and economic importance. Elucidating the genetic architecture underlying seed quality traits is essential for addressing persistent breeding challenges and improving the species' agronomic performance.

Sheepgrass demonstrates a broad distribution across eastern Eurasia, with disjunct populations found in North America and South Asia (Wu et al., 2018). This species exhibits remarkable ecological plasticity, showing exceptional tolerance to various abiotic stresses, including grazing pressure, soil salinity, and low temperatures (Chen et al., 2013). However, its agricultural potential is significantly constrained by reproductive limitations, notably self-incompatibility and seed dormancy mechanisms, which collectively result in poor seed set and low germination rates (Chen et al., 2019; Yang et al., 2019). These biological constraints present major obstacles to the species' commercial utilization, underscoring the need for targeted breeding interventions.

Conventional breeding strategies for sheepgrass, primarily relying on mass selection and hybridization techniques, face considerable challenges due to the species' complex allotetraploid genome (~8 Gb; Li et al., 2023) and obligate outcrossing nature (Chen et al., 2019). The inherent limitations of these traditional approaches—characterized by extended breeding cycles, intensive labor requirements, and modest genetic gains—highlight the urgent need for innovative solutions. Recent advancements in plant phenomics and genomics offer transformative potential, with high-throughput phenotyping platforms enabling precise trait quantification (Watt et al., 2020) and the recently completed reference genome sequence providing essential resources for molecular breeding applications.

The field of plant phenotyping has undergone revolutionary changes, evolving from basic manual measurements to sophisticated, automated systems capable of multidimensional trait analysis. Modern phenotyping platforms integrate cutting-edge imaging technologies (e.g., hyperspectral sensors, 3D laser scanning) with advanced

machine learning algorithms to extract comprehensive phenotypic data (Watt et al., 2020; Tang et al., 2023; Blanchy et al., 2025; Roth et al., 2025; Xu et al., 2025). While most phenotyping efforts have focused on vegetative structures, seed phenotyping has emerged as a distinct and increasingly important research domain. Traditional seed quality evaluation methods typically involve destructive sampling and subjective assessments (McDonald, 1998; Mahajan et al., 2015), whereas next-generation image-based platforms, such as AIseed (Tu et al., 2023), allow noninvasive, high-resolution characterization of morphological attributes—including size distribution, surface topography, and 3D architecture—with important applications in viability assessment and nutritional quality prediction (Mahajan et al., 2018; Wu et al., 2023; Qi et al., 2025).

Parallel developments in genotyping technologies have provided powerful tools for genetic analysis, with targeted sequencing approaches like Genotyping by Target Sequencing (GBTS) proving particularly advantageous for species with complex genomes. GBTS methodologies, including both multiplex PCR-based (GenoPlexs) and liquid-phase probe hybridization (GenoBaits) techniques, offer cost-efficient alternatives to whole-genome sequencing while maintaining high-density SNP detection capacity (Guo et al., 2021). These approaches have enabled significant achievements in genetic mapping, QTL analysis, and marker-assisted selection across numerous crop species, including wheat (*Triticum aestivum*) (Allen et al., 2017; Burridge et al., 2024), maize (*Zea mays*) (Ma et al., 2022), and soybean (*Glycine max*) (Liu et al., 2022). Notably, the successful application of GBTS in these agriculturally important species has demonstrated the technology's potential to accelerate genetic improvement in outcrossing perennial grasses.

Despite these advances, the integration of high-throughput phenotyping with advanced genotyping remains largely unexplored in sheepgrass research. To address this gap, we present a comprehensive study combining state-of-the-art seed phenotyping with SNP array genotyping to: (1) Utilizing the automated phenotyping workflow provided by the AIseed platform for the systematic characterization of sheepgrass seed traits; (2) Investigate phenotypic correlations between seed i-traits and agronomic performances, with a particular focus on germination-related traits; (3) Identify genomic regions associated with seed quality variation using a 50K SNP array, thereby establishing molecular tools for breeding applications. This integrative approach not only addresses critical limitations in sheepgrass improvement programs but also serves as a model for implementing phenomics-genomics integration in forage

crop breeding. The insights gained from this study will enhance our understanding of the genetic control of seed quality traits and provide practical tools for molecular breeding in this important species.

MATERIALS AND METHODS

Plant materials

In June 2024, at the Plant Germplasm Resource Garden of the Institute of Botany, Chinese Academy of Sciences (Beijing, N39°58', E116°12'), 262 hybrid progenies of sheepgrass were subjected to open pollination individually, and their offspring were harvested separately. These 262 offspring samples were subsequently used for further analyses.

Seed Phenotyping

From each sample, 100 seeds without glumes were randomly selected following ISTA sampling standards. Seeds were evenly spread on the glass surface of a flatbed scanner (HP OfficeJet Pro 8120, SN: 76AFEB), ensuring no overlap between seeds. Scanning was performed over an area of 160 × 240 mm at a resolution of 400 dpi, generating TIFF images. The AIseed software (v1.0; Tu et al., 2023) was employed to automatically extract 54 image-based traits (i-traits) including morphological parameters (e.g., length, width, perimeter, area), color features (e.g., RGB, Lab, HSV color spaces), and texture descriptors (e.g., entropy, contrast).

The hundred-seed weight (with and without glumes) was measured using a high-precision electronic balance (Meilen; accuracy ±0.001 g). For germination assays, 100 seeds without glumes were evenly placed on double-layered filter paper within 9-cm diameter plastic petri dishes. Subsequently, 10 mL of sterile deionized water was added, and dishes were incubated in a controlled environment chamber (Safe Intelligent Artificial Climate Chamber) under a 16 h light / 8 h dark photoperiod with day/night temperatures of 28℃/16℃ and 75% relative humidity. Water (5 mL) was supplemented every five days. Germination rate was assessed on the tenth day, defining germination as radicle length equal to or exceeding seed length.

Data preprocessing and statistical description

Trait data comprising 57 variables were analyzed using Python 3.7.6 to calculate descriptive statistics: minimum, maximum, range, mean, standard deviation (STD), skewness, kurtosis, coefficient of variation (CV), and Shannon-Wiener index (*H'*). Functional annotation of traits followed the classification criteria outlined by Mahajan

et al. (2018). All trait data were standardized using Z-scores ($z = (x - \mu)/\sigma$) to remove unit-scale effects prior to downstream analyses.

Correlation Analysis

Pearson's correlation coefficient (r) was used to assess linear relationships between standardized traits. Statistical significance was tested by two-tailed t-tests at a significance level of $\alpha = 0.05$. Correlations with $p < 0.05$ were considered significant (*), and those with $p < 0.01$ were considered highly significant (**). Correlation analyses were conducted using SPSS 25.0 (IBM Corp.), and correlation heatmaps were generated with OriginPro 2025 (OriginLab Corp.).

Elastic Net Regression Analysis

An elastic net regression model was constructed in Python 3.7.6 using scikit-learn 1.0.2. The dataset was randomly split into training (80%) and testing (20%) sets. Hyperparameters α (mixing coefficient for L1/L2 regularization) and λ (regularization strength) were optimized via 10-fold cross-validation. Traits with non-zero regression coefficients were identified as significant predictors. Trait weight maps were visualized using OriginPro 2025.

Principal Component Analysis

PCA was performed in SPSS 25.0. Components with eigenvalues greater than 1.0 and a cumulative variance contribution of at least 95% were retained. Both the variable loading matrix and sample score matrix were recorded. The comprehensive score (F) for each sample was calculated as: $F = \sum_{i=1}^{k} (PC_i \times \omega_i)$, where k is the number of principal components, PC_i represents the sample score on the i-th component, and ω_i denotes the variance contribution weight of the i-th component.

DNA extraction and genotyping

Genomic DNA was extracted from 100 mg of fresh young leaves using the CTAB method (Doyle et al., 1987). DNA quality was assessed by electrophoresis on 1% agarose gels. Concentration and purity were measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher). Genotyping was performed using a custom 50K SNP array developed for sheepgrass. Quality control was performed by excluding SNP loci with missing rates greater than 10%, imputing missing genotypes using the k-nearest neighbor algorithm ($k = 10$), and filtering out loci with a minor allele frequency (MAF) less than 0.05.

Genome-wide association analysis

GWAS were conducted on 54 i-traits using the linear mixed model (LMM)

implemented in GEMMA v0.98.1 (Zhou et al., 2012). The kinship matrix was computed with the command "-gk 1", and association analysis was performed using "-lmm 4". The genome-wide significance threshold was set at $p < 1 \times 10^{-5}$. Manhattan and Q-Q plots were generated using R 4.1.0 with the qqman package.

Candidate gene screening and functional enrichment

Significant SNPs ($p < 1 \times 10^{-5}$) were mapped to the sheepgrass reference genome (Li et al., 2023). Genes containing or adjacent to these SNPs were identified as candidate genes. KEGG pathway enrichment analysis was performed using KOBAS 3.0 (Bu et al., 2021), with multiple testing correction via the Benjamini-Hochberg method and a false discovery rate (FDR) threshold of ≤ 0.05. Functional annotation of candidate genes was conducted using the ANNOVAR database (Wang et al., 2010).

RESULTS

Basic statistical analysis of sheepgrass seed traits

In this study, a total of 262 sheepgrass samples were scanned, and 54 i-traits were extracted using AIseed. These included morphological traits (e.g., length [L], width [W], length-to-width ratio [L/W Ratio], area [A], perimeter [P], roundness [O]), color traits (statistical indicators in RGB and Lab color spaces), and texture traits (measured based on the correlation of the grayscale co-occurrence matrix and the RGB channels). Additionally, three agronomic traits—hundred grain weight (HGW), hundred grain weight without glumes (HGWwg), and germination rate (GR)—were manually measured. The statistical results are summarized in Table 1.

Morphologically, sheepgrass seeds exhibited a slender shape, with an average L/W Ratio of 2.69 ± 0.12, a perimeter of 10.13 ± 0.42 mm, and an area of 3.86 ± 0.24 mm².

Analysis of color traits showed that the mean RGB values of the seed coat were R=114.18 ± 10.89, G=133.06 ± 9.32, and B=132.00 ± 6.38, with standard deviations exceeding 38 across all channels, indicating substantial genetic diversity in seed coloration. Visual inspection revealed a range of seed coat colors, including yellow, reddish-brown, and dark brown, which are likely associated with genetic background and maturity. The mean grayscale value was 127.29 ± 9.35, with a considerable standard deviation (40.00 ± 2.85), suggesting significant variation in lightness and darkness possibly due to differences in seed coat texture and pigment distribution.

Texture analysis demonstrated that the seed surface had high gray contrast (Gray_contrast: 1074.45 ± 67.22) and dissimilarity (Gray_dissimilarity: 11.33 ± 0.66),

along with high homogeneity (Gray_homogeneity: 0.70 ± 0.01), energy (Gray_energy: 0.37 ± 0.02), and correlation (Gray_correlation: 0.88 ± 0.01). These results indicate a uniform and linearly arranged texture distribution along the longitudinal axis. Moreover, the high entropy value (Gray_entropy: 4.92 ± 0.11) suggests a complex microstructure of the seed coat, potentially related to features such as hairs, cracks, or a waxy layer.

Agronomic traits analysis showed that HGW (0.27 ± 0.03 g) was higher than HGWwg (0.15 ± 0.03 g), with glumes' weight accounting for approximately 44.44% of the total. GR varied substantially, ranging from 0.07 to 0.90, with a mean of 0.47 ± 0.14, indicating significant differences in germination capacity among the samples.

The statistical analysis of all traits demonstrated that skewness and kurtosis values met the basic assumptions for normality (| Skewness | < 3, | Kurtosis | < 10), supporting the use of subsequent analyses such as Pearson's correlation, linear regression, PCA, and GWAS. The CV for the traits ranged from 1.23% to 30.44%, with an average of 6.86%. Notably, GR (CV = 30.44%), the mean and standard deviation of the Lab b channel (b_mean, CV = 23.38%; b_std, CV = 20.00%), HGWwg (CV = 19.74%), and the STD and mean of the Lab a channel (a_std, CV = 12.93%; a_mean, CV = 12.15%) exhibited considerable variation, while the CVs of the remaining traits were all below 10%. The average Shannon-Wiener index (*H'*) was 2.064. Except for HGW (*H'* = 1.983) and HGWwg (*H'* = 1.992), all other traits had *H'*-values above 2.000, reflecting rich phenotypic diversity and uniform distribution within the population.

Table 1 Basic Statistical Analysis of *Leymus chinensis* Seed Traits

Character classification	Abbreviation	Trait definition	Minimum	Maximum	Range	Mean	STD	Skewness	Kurtosis	CV	*H'*
Morphological traits	L (mm)	Length	3.34	4.28	0.94	3.79	0.18	0.084	-0.056	4.65%	2.072
	W (mm)	Width	1.24	1.51	0.27	1.42	0.04	-0.177	0.678	2.92%	2.064
	L/W Ratio	Length/Width Ratio	2.38	3.16	0.78	2.69	0.12	0.365	0.547	4.62%	2.074
	A (mm^2)	Area	3.28	4.52	1.24	3.86	0.24	0.058	-0.284	6.25%	2.097
	P (mm)	Perimeter	9.11	11.32	2.21	10.13	0.42	0.090	-0.067	4.13%	2.078
	O	Roundness	0.42	0.52	0.10	0.47	0.02	-0.140	0.389	3.22%	2.059
Color traits	R_mean	Mean of Red Value	89.50	141.80	52.29	114.18	10.89	0.307	-0.596	9.54%	2.075
	R_std	Standard Deviation of Red Value	31.09	44.03	12.94	38.17	2.23	-0.174	0.029	5.84%	2.082
	G_mean	Mean of Green Value	111.64	158.34	46.70	133.06	9.32	0.365	-0.226	7.00%	2.064
	G_std	Standard Deviation of Green Value	34.78	52.80	18.01	42.35	3.09	0.216	-0.242	7.29%	2.082
	B_mean	Mean of Blue Value	117.43	153.62	36.19	132.00	6.38	0.508	0.382	4.83%	2.043

	B_std	Standard Deviation of Blue Value	39.55	55.24	15.69	45.92	2.86	0.344	-0.335	6.22%	2.063
	l_mean	Mean of Lightness Value	44.93	63.39	18.45	53.71	3.70	0.306	-0.333	6.90%	2.062
	l_std	Standard Deviation of Lightness Value	12.87	20.29	7.42	16.02	1.28	0.222	-0.277	7.99%	2.082
	a_mean	Mean of Red-Green Value	1.57	3.36	1.79	2.36	0.29	0.309	0.301	12.15%	2.061
	a_std	Standard Deviation of Red-Green Value	1.01	2.32	1.31	1.62	0.21	0.148	0.243	12.93%	2.040
	b_mean	Mean of Yellow-Blue Value	4.27	12.98	8.71	7.94	1.86	0.271	-0.715	23.38%	2.066
	b_std	Standard Deviation of Yellow-Blue Value	2.98	7.58	4.59	4.91	0.98	0.231	-0.712	20.00%	2.065
	H_mean	Mean of Hue Value	58.79	88.62	29.83	73.33	6.34	0.056	-0.810	8.65%	2.063
	H_std	Standard Deviation of Hue Value	28.06	36.12	8.06	32.32	1.49	-0.258	0.024	4.62%	2.068
	S_mean	Mean of Saturation	55.50	77.35	21.85	64.01	3.73	0.458	0.046	5.83%	2.052
	S_std	Standard Deviation of Saturation	29.56	45.64	16.08	35.92	3.40	0.353	-0.664	9.46%	2.056
	V_mean	Mean of Brightness	121.75	164.69	42.94	140.96	8.49	0.424	-0.183	6.02%	2.049
	V_std	Standard Deviation of Brightness	34.10	53.77	19.67	42.31	3.63	0.247	-0.439	8.59%	2.082
	Gray_mean	Mean of Grayscale Value	105.67	152.29	46.62	127.29	9.35	0.356	-0.310	7.34%	2.055
	Gray_std	Standard Deviation of Grayscale Value	32.39	49.51	17.12	40.00	2.85	0.123	-0.183	7.12%	2.091
Textural traits	Gray_contrast	Contrast of Grayscale	929.71	1336.92	407.21	1074.45	67.22	0.469	0.253	6.26%	2.048
	Gray_dissimilarity	Dissimilarity of Grayscale	9.56	13.98	4.42	11.33	0.66	0.331	0.296	5.87%	2.068
	Gray_homogeneity	Homogeneity of Grayscale	0.67	0.73	0.06	0.70	0.01	0.185	0.224	1.25%	2.053
	Gray_energy	Energy of Grayscale	0.33	0.44	0.12	0.37	0.02	0.171	0.332	4.78%	2.062
	Gray_correlation	Correlation of Grayscale	0.84	0.91	0.07	0.88	0.01	-0.015	0.157	1.46%	2.093
	Gray_ASM	Angular Second Moment of Grayscale	0.11	0.21	0.10	0.15	0.01	0.264	0.568	9.20%	2.052
	Gray_entropy	Entropy of Grayscale	4.57	5.16	0.59	4.92	0.11	-0.253	0.018	2.21%	2.078
	R_contrast	Contrast of Red Channel	623.45	925.07	301.63	735.42	54.14	0.473	0.001	7.36%	2.051
	R_dissimilarity	Dissimilarity of Red Channel	8.63	12.22	3.59	10.02	0.58	0.309	0.047	5.78%	2.063
	R_homogeneity	Homogeneity of Red Channel	0.67	0.73	0.06	0.70	0.01	0.131	0.267	1.23%	2.076
	R_energy	Energy of Red Channel	0.33	0.44	0.12	0.37	0.02	0.169	0.337	4.79%	2.063
	R_correlation	Correlation of Red Channel	0.85	0.93	0.08	0.90	0.01	-0.257	0.028	1.54%	2.071
	R_ASM	Angular Second Moment of Red Channel	0.11	0.21	0.10	0.15	0.01	0.264	0.574	9.20%	2.053

	R_entropy	Entropy of Red Channel	4.54	5.14	0.60	4.90	0.10	-0.258	0.124	2.13%	2.056
	G_contrast	Contrast of Green Channel	1056.24	1509.92	453.68	1222.36	73.85	0.446	0.256	6.04%	2.063
	G_dissimilarity	Dissimilarity of Green Channel	10.10	14.67	4.57	11.91	0.69	0.325	0.317	5.79%	2.078
	G_homogeneity	Homogeneity of Green Channel	0.67	0.73	0.06	0.70	0.01	0.188	0.211	1.26%	2.057
	G_energy	Energy of Green Channel	0.33	0.44	0.12	0.37	0.02	0.167	0.332	4.79%	2.062
	G_correlation	Correlation of Green Channel	0.84	0.91	0.07	0.88	0.01	0.035	0.189	1.41%	2.078
	G_ASM	Angular Second Moment of Green Channel	0.11	0.21	0.10	0.15	0.01	0.259	0.577	9.20%	2.050
	G_entropy	Entropy of Green Channel	4.59	5.19	0.60	4.95	0.11	-0.261	0.042	2.21%	2.074
	B_contrast	Contrast of Blue Channel	4.59	5.19	0.60	4.95	0.11	-0.261	0.042	2.21%	2.074
	B_dissimilarity	Dissimilarity of Blue Channel	11.30	16.03	4.73	13.08	0.75	0.300	0.316	5.70%	2.096
	B_homogeneity	Homogeneity of Blue Channel	0.67	0.73	0.06	0.69	0.01	0.158	0.259	1.26%	2.085
	B_energy	Energy of Blue Channel	0.33	0.44	0.12	0.37	0.02	0.165	0.334	4.80%	2.062
	B_correlation	Correlation of Blue Channel	0.82	0.89	0.07	0.86	0.01	0.158	0.283	1.30%	2.072
	B_ASM	Angular Second Moment of Blue Channel	0.11	0.21	0.10	0.15	0.01	0.257	0.577	9.21%	2.056
	B_entropy	Entropy of Blue Channel	4.62	5.25	0.64	5.00	0.11	-0.289	0.213	2.17%	2.081
Agronomic traits	HGW (g)	Hundred Grain Weight	0.14	0.39	0.25	0.27	0.03	-0.108	0.869	12.81%	1.983
	HGWwg (g)	Hundred Grain Weight without glumes	0.05	0.23	0.18	0.15	0.03	0.047	0.080	19.74%	1.992
	GR	Germination Rate	0.07	0.90	0.83	0.47	0.14	0.141	-0.443	30.44%	2.068

Pearson's correlation analysis of seed traits

Pearson correlation analysis was conducted to explore the relationships among 57 phenotypic traits of sheepgrass seeds based on data from 262 samples. The resulting correlation coefficients (r) and significance levels (p) are summarized in the heatmap presented in Figure 1. Significant positive correlations ($p < 0.01$) were observed among the six morphological traits, except for O, which exhibited strong negative correlations with other morphological traits ($r = -0.96$ to -0.45). The L/W Ratio showed a modest negative correlation with W ($r = -0.30$), while O showed a modest positive correlation with W ($r = 0.26$). Morphological traits generally displayed weak correlations with color and agronomic traits (HGW and HGWwg), with most $|r| < 0.50$. Notably, W was

moderately and positively correlated with HGW ($r = 0.52$, $p < 0.01$) and HGWwg ($r = 0.47$, $p < 0.01$). The relationships between morphological and textural traits were complex: L, L/W Ratio, and P correlated moderately to strongly with multiple textural traits ($r = 0.59$ to 0.78, $p < 0.01$), whereas O correlated negatively with most textural traits ($r = -0.91$ to -0.85, $p < 0.01$). Although several texture indices (e.g., Gray_contrast, Gray_dissimilarity, Gray_correlation) reached statistical significance at the 0.05 level, their correlation strengths were mostly weak ($|r| < 0.50$).

Strong correlations were detected among the 20 color traits ($r = -0.91$ to 1.00, $p < 0.01$), except for a mean, a std, and H std, which exhibited weaker correlations with other color traits (most $|r| < 0.40$) but were strongly correlated among themselves ($r = 0.67$ to 0.98, $p < 0.01$). Correlations between color traits and both texture and agronomic traits were generally weak and nonsignificant ($|r| < 0.30$), except for texture "correlation" indices (Gray_correlation, R correlation, G correlation, B correlation), which displayed moderate to strong correlations with several color traits ($r = -0.80$ to 0.85, $p < 0.01$). Within textural traits, "correlation" variables showed near-perfect inter-correlations ($r \approx 1.00$, $p < 0.01$) and strong positive correlations with "dissimilarity" indices ($|r| = 0.70$ to 0.99, $p < 0.01$). Some contrast and correlation indices, while statistically significant ($p < 0.01$), exhibited relatively weak correlations (most $|r| < 0.60$). Correlations between textural and agronomic traits (HGW, HGWwg) were negligible ($|r| < 0.20$, $p > 0.05$), indicating limited direct predictive potential of texture for yield traits. Among agronomic traits, HGW and HGWwg were moderately and significantly correlated ($r = 0.48$, $p < 0.01$), whereas GR showed negligible and nonsignificant correlations with both HGW and HGWwg ($r < 0.20$, $p > 0.05$).

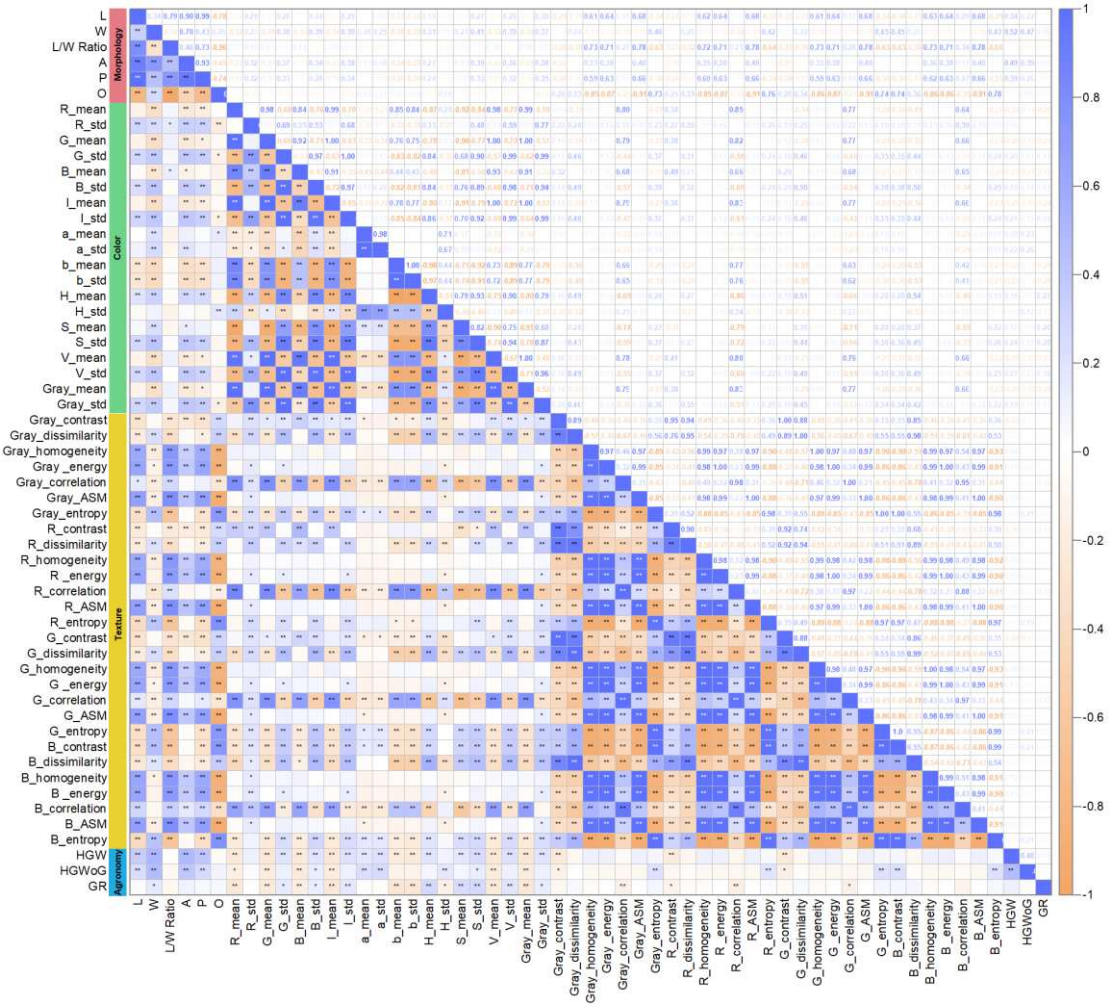

Figure 1 Correlation analysis of *Leymus chinensis* seed traits. The upper right corner shows the correlation coefficients (*r*), while the lower left corner displays the significance levels of the correlations. The size of the squares and the depth of their colors correspond to the magnitude of the correlation coefficients. Significance is determined at $p < 0.05$ (*) and high significance at $p < 0.01$ (**). The color bar on the right indicates the size of the correlation coefficients, with blue representing positive correlations and orange representing negative correlations. Character classifications are labeled on the left side of the figure: Morphology, Color, Texture, and Agronomy.

Elastic network regression analysis of seed i-traits and agronomic traits

The above correlation analysis (Figure 1) revealed substantial multicollinearity among the seed i-traits, with maximum correlation coefficients reaching $|r| = 1.00$. Such high multicollinearity reduces the explanatory power of individual variables in traditional regression models and increases the risk of overfitting. To address these challenges and effectively identify key explanatory variables from high-dimensional, collinear data, we applied an elastic net regression model. This approach integrates L1 regularization (Lasso), which performs variable selection, with L2 regularization (Ridge), which

mitigates multicollinearity and stabilizes coefficient estimates. The elastic net thus provides a robust theoretical framework for screening candidate traits for high-yield breeding indices.

For HGW, optimal hyperparameters were determined via 10-fold cross-validation as $\lambda = 0.074$ and $\alpha = 0.73$. The relatively small λ indicates modest shrinkage of coefficients, while $\alpha > 0.5$ suggests a model primarily driven by L1 regularization, favoring variable selection. Under these settings, five traits were retained with non-zero standardized regression coefficients (Figure 2A). Among these, W ($\beta = 0.349$) and A ($\beta = 0.169$) exhibited positive contributions to HGW, whereas the textural trait gray-level contrast (G_contrast; $\beta = -0.124$) showed a negative association. Gray homogeneity (Gray_homogeneity; $\beta = 0.019$) and green mean intensity (G_mean; $\beta = -0.013$) had minimal influence.

A similar cross-validation approach for HGWwg yielded optimal hyperparameters $\lambda = 0.002$ and $\alpha = 0.75$. The extremely low λ implies very weak coefficient penalization, while the high α again reflects a model dominated by L1 regularization, enabling extensive variable selection. Consequently, 26 traits were selected with non-zero coefficients (Figure 2B), encompassing the key traits of HGW (W, A, G_contrast, G_mean, Gray_homogeneity). Notably, textural traits exhibited the strongest effects, with absolute coefficients ranging from 0.930 to 1.656, followed by color traits. Morphological traits had comparatively smaller contributions, with the largest morphological coefficient corresponding to P ($|\beta| = 0.311$). The markedly greater number of important variables for HGWwg relative to HGW underscores the enhanced predictive relevance and complexity of texture-related traits for this agronomic measure.

For GR, cross-validation identified $\lambda = 0.001$ and $\alpha = 0.99$, indicating an essentially pure Lasso model. This model selected thirteen important traits from the initial 54 i-traits (Figure 2C), including four core traits (W, S_mean, S_std, B_std) common to both HGW and HGWwg models. The top four predictors, ranked by absolute coefficient magnitude, were color traits, highlighting their dominant explanatory power for GR. Textural traits followed with moderate influence (mean $|\beta| = 0.026$), while morphological traits had the weakest effects (mean $|\beta| = 0.009$). These findings align with previous research demonstrating the critical role of seed coat pigmentation in seed dormancy and germination processes (Finkelstein et al., 2008; Nonogaki, 2019).

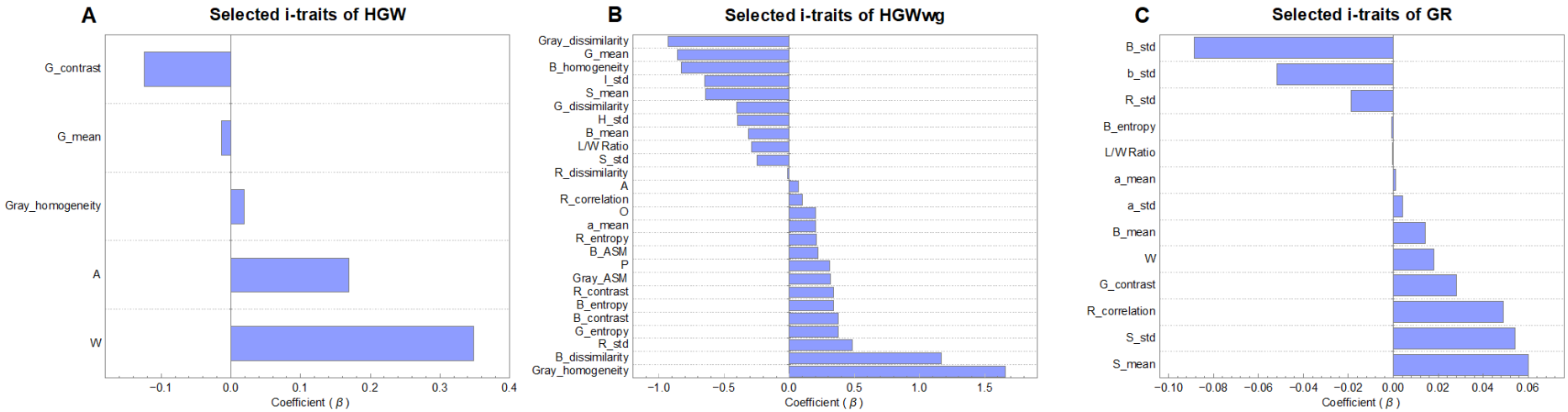

Figure 2 Key i-traits linked to agronomic traits and their coefficient distribution, as screened by elastic net. Sections A, B, and C respectively show important image-based traits (i-traits) for hundred-seed weight (HGW), hundred-seed weight without glumes (HGWwg), and germination rate (GR). Bar length reflects the magnitude of standardized regression coefficients (β), with +/− indicating positive/negative.

Principal component analysis of seed i-traits

PCA was conducted on 54 seed i-traits, resulting in the extraction of six principal components (PC1 to PC6) with eigenvalues greater than 1. These components accounted for a cumulative variance of 96.32% (Table 2), indicating effective dimensionality reduction. The first three principal components explained 40.57%, 29.30%, and 11.30% of the total variance, respectively.

Variables with loadings of 0.7 or higher were considered strong contributors to each principal component (Table 3). PC1 (Texture Homogeneity Factor) contributed 40.57% of the total variance and was mainly associated with gray/R/G/B homogeneity, energy, and ASM (positive loadings), as well as entropy (negative loadings), characterizing the regularity and uniformity of seed texture. PC2 (Color Intensity and Shape Size Factor), accounting for 29.30% of the variance, was primarily defined by color variables such as H_mean, S_std, and B_mean, along with shape indicators including P, A, and L, reflecting the positive relationship between color attributes and seed size. PC3 (Texture Contrast and Color Variability Factor) explained 11.30% of the variance, with contributions mainly from gray/R/G contrast, B_mean, and R_std, representing local texture contrast and color channel variability. PC4 (Morphological Width and Color Variability Factor) contributed 7.55%, primarily influenced by W, R_std, and B_correlation, reflecting morphological width and color variability. PC5 (Lab Color Space Trait Factor) accounted for 4.85% of the variance, defined by a_mean, a_std, and H_std, capturing variation in red and green color channels. PC6, contributing 3.03%, had no variables with significant loadings (absolute values < 0.5) and lacked clear biological interpretation. Overall, texture traits were mainly represented in PC1 and PC3, color traits in PCs 2, 3, and 5, and morphological traits in PCs 2 and 4.

Comprehensive scores (F) for each germplasm line were calculated based on principal component contribution rates and individual sample scores, with weights assigned according to the correlation between principal components and agronomic traits. PC1, PC2, and PC4 received positive weights, while PC3 and PC5 were weighted negatively. Germplasm lines were ranked by their comprehensive scores, and the population mean was used to define “high-quality germplasm.” Morphological quality was assessed with L and W as positive indicators; color quality was assessed with S_mean as a positive and V_mean as a negative indicator; texture quality was assessed with Gray_contrast as a negative indicator; and agronomic performance was evaluated

using HGW, HGWwg, and GR as positive indicators. The top ten germplasm lines in the comprehensive ranking demonstrated superior traits, including large seed size, dark coloration, high fullness, elevated HGW, and high GR (Table 4).

Furthermore, PC2 and PC3 showed strong correlations with germination rate: PC2 was positively correlated, whereas PC3 was negatively correlated. Accordingly, positive and negative weights were assigned to PC2 and PC3, respectively, in the scoring system. Using a mean germination rate of 0.47 as the threshold for high germination, the top 45 samples were selected based on their comprehensive scores. Among these, 32 samples exhibited actual GRs above the mean, resulting in a prediction accuracy of 71%. The binomial test yielded $p = 0.003$ (95% CI: 56.6% to 82.3%; $p < 0.05$), which was significantly higher than random selection. These results indicate that the scoring system effectively identified seed lines with high germination potential.

Table 2 Eigenvalues and Contribution Rates of Principal Components of *Leymus Chinensis* Seed i-Traits

Principal component	Eigenvalue	Contribution (%)	Cumulative contribution (%)
PC1	21.906	40.57%	40.57%
PC2	15.675	29.03%	69.59%
PC3	6.103	11.30%	80.90%
PC4	4.076	7.55%	88.44%
PC5	2.618	4.85%	93.29%
PC6	1.637	3.03%	96.32%

Table 3 Principal Component Loading Matrix of *Leymus Chinensis* Seed i-Traits

Traits	Principal component					
	PC1	PC2	PC3	PC4	PC5	PC6
L	0.491	0.539	0.137	0.367	0.341	-0.412
W	-0.324	0.303	-0.061	0.557	0.287	-0.309
L/W Ratio	0.697	0.351	0.172	0.018	0.155	-0.227
A	0.194	0.517	0.070	0.551	0.395	-0.467
P	0.448	0.579	0.134	0.395	0.340	-0.410
O	-0.760	-0.472	-0.208	0.051	-0.107	0.132
R_mean	0.418	-0.828	0.321	0.139	0.103	0.036
R_std	-0.042	0.311	0.628	0.586	-0.133	0.293
G_mean	0.417	-0.775	0.451	0.124	-0.002	0.016
G_std	-0.347	0.796	0.254	0.308	-0.091	0.258
B_mean	0.327	-0.562	0.697	0.174	-0.078	0.028
B_std	-0.420	0.808	0.145	0.201	-0.012	0.235

l_mean	0.417	-0.791	0.425	0.121	0.021	0.015
l_std	-0.356	0.811	0.239	0.299	-0.104	0.225
a_mean	-0.161	0.092	-0.493	0.095	0.707	0.419
a_std	-0.157	0.138	-0.443	0.141	0.728	0.403
b_mean	0.387	-0.845	-0.121	-0.024	0.138	0.039
b_std	0.394	-0.835	-0.120	-0.039	0.141	0.055
H_mean	-0.402	0.854	0.105	-0.040	-0.196	0.000
H_std	0.094	-0.363	-0.456	0.220	0.621	0.281
S_mean	-0.415	0.763	-0.293	-0.125	-0.190	0.059
S_std	-0.399	0.853	0.075	0.185	-0.172	0.028
V_mean	0.398	-0.761	0.481	0.139	0.013	0.036
V_std	-0.393	0.842	0.200	0.213	-0.108	0.180
Gray_mean	0.415	-0.786	0.430	0.135	0.029	0.024
Gray_std	-0.313	0.751	0.341	0.365	-0.107	0.262
Gray_contrast	-0.533	-0.113	0.726	-0.332	0.225	0.049
Gray_dissimilarity	-0.733	0.203	0.561	-0.226	0.197	-0.021
Gray_homogeneity	0.920	0.368	0.041	-0.040	0.008	0.070
Gray _energy	0.836	0.504	0.135	-0.089	0.044	0.063
Gray_correlation	0.704	-0.505	0.114	0.442	-0.129	0.117
Gray_ASM	0.843	0.500	0.147	-0.095	0.064	0.017
Gray_entropy	-0.913	-0.083	-0.004	0.387	-0.037	-0.044
R_contrast	-0.395	-0.299	0.766	-0.193	0.326	0.066
R_dissimilarity	-0.641	-0.007	0.687	-0.138	0.267	-0.008
R_homogeneity	0.891	0.434	0.012	-0.054	0.003	0.065
R_energy	0.834	0.507	0.134	-0.090	0.044	0.062
R_correlation	0.675	-0.624	0.027	0.358	-0.075	0.089
R_ASM	0.841	0.502	0.146	-0.095	0.065	0.017
R_entropy	-0.854	-0.250	0.045	0.440	-0.030	-0.033
G_contrast	-0.545	-0.097	0.707	-0.370	0.192	0.044
G_dissimilarity	-0.745	0.216	0.531	-0.258	0.186	-0.024
G_homogeneity	0.917	0.376	0.048	-0.034	0.006	0.070
G_energy	0.836	0.504	0.135	-0.089	0.042	0.063
G_correlation	0.709	-0.469	0.128	0.461	-0.149	0.124
G_ASM	0.843	0.500	0.146	-0.096	0.061	0.016
G_entropy	-0.919	-0.085	-0.034	0.367	-0.024	-0.054
B_contrast	-0.919	-0.085	-0.034	0.367	-0.024	-0.054
B_dissimilarity	-0.774	0.236	0.441	-0.331	0.168	-0.039
B_homogeneity	0.885	0.444	0.068	-0.021	0.007	0.084
B_energy	0.834	0.507	0.134	-0.089	0.040	0.063
B_correlation	0.695	-0.273	0.203	0.567	-0.197	0.152
B_ASM	0.842	0.502	0.145	-0.096	0.060	0.016
B_entropy	-0.918	-0.178	-0.066	0.317	0.012	-0.096

Table 4 Top 10 of *Leymus Chinensis* Germplasms' Comprehensive Scores and Rankings

Comprehensive excellent germplasm		Reference seed i-traits					Reference Agronomic traits		
No.	*F*	L	W	S_mean	V_mean	Gray_contrast	HGW	HGWwg	GR
223	8.808	4.28	1.40	72.40	129.09	1025.44	0.32	0.18	0.41
190	5.038	4.13	1.43	66.58	130.78	987.44	0.31	0.15	0.65
12	5.002	4.05	1.45	62.14	138.65	971.43	0.33	0.18	0.66
150	4.853	4.22	1.40	61.49	148.74	997.53	0.23	0.12	0.63
493	4.778	3.89	1.24	57.74	164.20	1042.68	0.18	0.05	0.72
79	4.676	4.08	1.39	63.26	135.56	943.49	0.28	0.15	0.49
221	4.565	3.99	1.39	67.37	133.57	945.29	0.23	0.15	0.56
393	4.448	4.11	1.45	67.73	128.13	975.42	0.29	0.15	0.55
490	4.300	3.81	1.41	64.82	138.82	945.21	0.28	0.14	0.51
490	3.986	4.15	1.47	66.02	131.97	964.35	0.31	0.2	0.45
Mean		3.79	1.42	64.01	140.96	1074.45	0.27	0.15	0.47

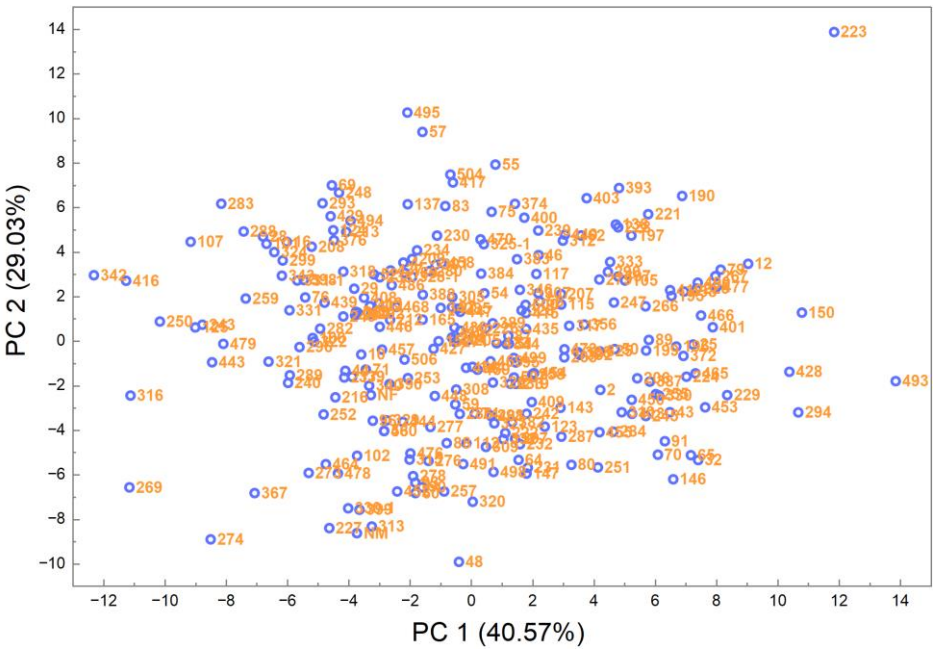

Figure 3 PCA scatter plot of 262 *Leymus chinensis* germplasms. The scatter plot of PC1 versus PC2 shows continuous distribution, indicating no significant subpopulation differentiation within the study population.

Genome-wide association analysis of seed i-traits

In this study, a self-developed sheepgrass 50K SNP array was employed to perform high-throughput genotyping on 262 sheepgrass germplasm samples. The objective was to elucidate the molecular basis underlying the correlations between i-traits and Agronomic traits, and to identify genetic loci influencing phenotypic variation in sheepgrass seeds. A GWAS was conducted on 54 i-traits to detect significantly

associated loci.

Multiple SNPs significantly associated with color traits—namely R, G, B, H, S, V, b, and Gray—were identified. These SNPs were predominantly distributed on chromosome Lc2Xm, with some loci overlapping across different color traits. SNPs significantly associated with the morphological trait L were mainly located on chromosome Lc6Xm (Figure 4). Functional annotations of the most significant SNP loci for each trait are summarized in Table 5.

The top SNP locus associated with R, S, V, G, and Gray traits (Lc2Xm: 433,045,629 bp) is positioned 772 bp upstream of the transcription start site of a gene homologous to *At5g07610* in *Arabidopsis thaliana*, with a MAF of 0.47 (G to A). The most significant SNPs for b and H traits are located at Lc2Xm: 424,242,352 bp and Lc2Xm: 424,243,081 bp, respectively. Both reside in a noncoding intergenic region between unannotated genes Lc2Xm053490 and Lc2Xm053491, separated by 1.3 to 2.0 kb. The top SNP associated with B (Lc2Xm: 438,932,789 bp) is located 579 bp downstream of the 3' end of the *Expansin B4* (*EXPB4*) gene. The leading SNP linked to length (L) (Lc6Xm: 345,082,398 bp) is situated within an intron of the *Xap5 circadian timekeeper* (*XCT*) gene.

KEGG pathway enrichment analysis was performed on genomic segments harboring significant SNPs from nine traits, revealing nine pathways significantly enriched (q-value < 0.05; Figure 5). These pathways involved 27 transcripts corresponding to 26 annotated genes.

Terpenoid and Flavonoid Biosynthesis Pathways: Monoterpenoid biosynthesis was significantly enriched in six trait-associated segments (R, G, H, V, B, and Gray), all of which contained *Salutaridine reductase* (*SALR*) and (+)-*Neomenthol dehydrogenase* (*MNR1*) genes separated by approximately 8.4 kb. This suggests the presence of a *Short-chain dehydrogenase/reductase* (*SDR*) gene cluster on chromosome Lc2Xm spanning 395.496 to 395.504 Mb. Diterpenoid biosynthesis was enriched in R, S, and B segments, each containing *Acyclic sesquiterpene synthase* (*TPS1*) and a *TPS1*-like gene separated by 15 kb, indicating a *TPS* gene cluster at Lc2Xm 434.873 to 434.917 Mb. Flavonoid biosynthesis was enriched in R, G, V, and Gray regions, which harbor *Anthocyanidin reductase* (*ANR*) and two *ANR*-like genes within less than 5 kb, suggesting an *ANR* gene cluster in Lc2Xm 478.150 to 478.199 Mb.

Epidermal Structural Metabolism Pathways: Cutin, suberin, and wax biosynthesis pathways were significantly enriched in seven trait-associated segments,

excluding L and B. The S segment contained three *Peroxygenase*-like (*PXG*-like) genes, while the other segments contained peroxygenase (*PXG*) and multiple *PXG*-like genes. Most gene intervals were under 5 kb, indicative of a *PXG* gene cluster located on Lc2Xm between 394.244 and 394.408 Mb.

Amino Acid Metabolism Pathways: Alanine, aspartate, and glutamate metabolism pathways were significantly enriched in G, Gray, R, and V segments. These segments include genes encoding *Gamma-aminobutyrate transaminase 1, mitochondrial* (*OsI_17385*), *Chloroplastic glutamine synthetase leaf isozyme* (*GLNA*), *Mitochondrial aspartate aminotransferase* (*ASP1*), and *MNR1*. The arginine and proline metabolism pathway was enriched specifically in the R segment, containing *S-adenosylmethionine decarboxylase proenzyme* (*SAMDC*), *Aldehyde Dehydrogenase 3 Family Member F1* (*ALDH3F1*), and *ASP1*.

Signal Transduction and Other Pathways: The plant MAPK signaling pathway was enriched in the B and R regions, incorporating genes such as *1-Aminocyclopropane-1-Carboxylate Synthase* (*ACS1*), *Heavy Metal ATPase HMA5 (HMA5)* and *HMA5*-like, and *Mitogen-Activated Protein Kinase Kinase Kinase YODA* (*YDA*). A 7.4 kb distance separates *HMA5* and *HMA5*-like, suggesting an *HMA* gene cluster within Lc2Xm between 424.716 and 424.729 Mb. Phenylpropanoid biosynthesis was exclusively enriched in the R region, encompassing *Peroxidase 18* (*PER18*), *Beta-glucosidase 16* (*BGLU16*), *Phenylalanine ammonia-lyase* (*PAL*), *Beta-glucosidase BoGH3B* (*BACOVA_02659*), and *Peroxidase 50* (*PER50*). The L segment displayed exclusive enrichment in the mRNA surveillance pathway, with particular emphasis on the *RNA guanylyltransferase and 5'-phosphatase* (*Rngtt*) gene. Notably, no significant pathways related to pigment synthesis or epidermal structure were detected in this segment.

Table 5 The Most Significant SNPs Information

Traits	SNP Position (bp)	Candidate gene	Position of SNP relative to the candidate gene	MAF
R/G/S/V/Gray	Lc2Xm: 433045629	*At5g07610*	Upstream (dist = 772 bp)	0.47
b	Lc2Xm: 424242352	*Lc2Xm053490*	Intergenic (*Lc2Xm053490* (dist = 1312 bp), *Lc2Xm053491* (dist = 149573 bp))	0.40
H	Lc2Xm: 424243081	*Lc2Xm053490*	Intergenic (*Lc2Xm053490* (dist = 2041 bp), *Lc2Xm053491* (dist = 148844 bp))	0.44
B	Lc2Xm: 438932789	*EXPB4*	Downstream (dist = 579 bp)	0.49
L	Lc6Xm: 345082398	*XCT*	Intronic	0.19

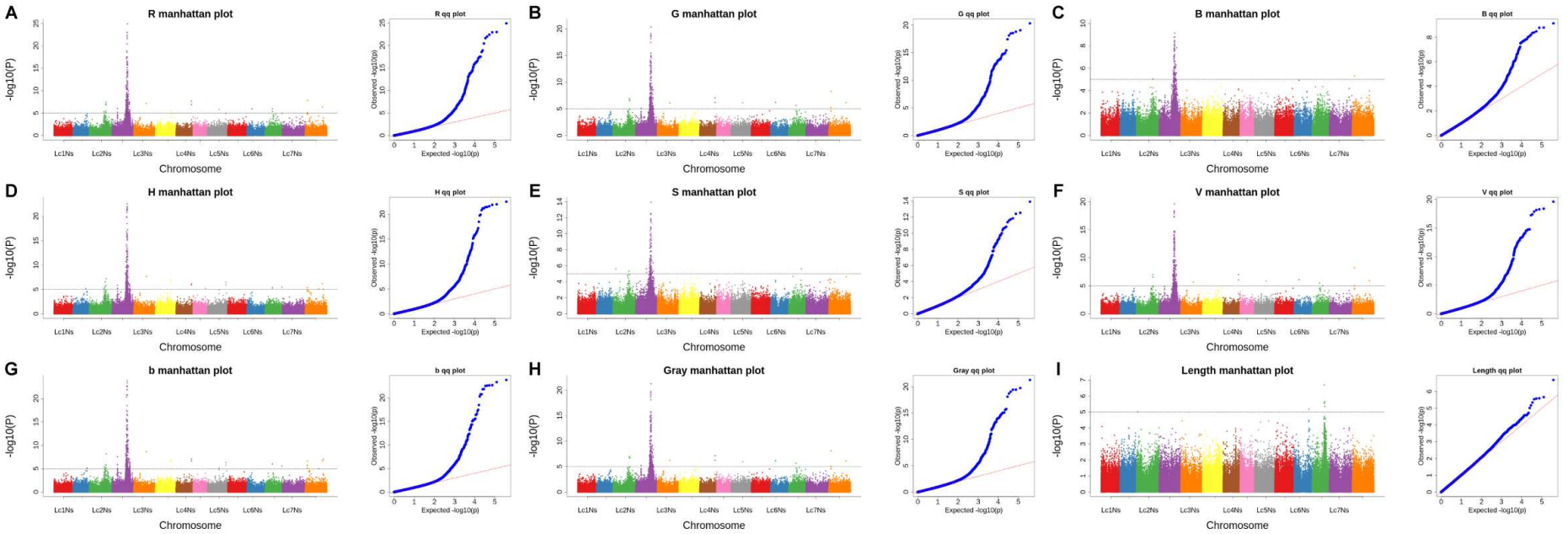

Figure 4 Manhattan and Q-Q plots for R, G, B, H, S, V, b, Gray and Length. Here, the color traits correspond to R_mean, G_mean, B_mean, H_mean, S_mean, V_mean, b_mean, and Gray_mean mentioned earlier; the morphological trait correspond to L mentioned earlier.

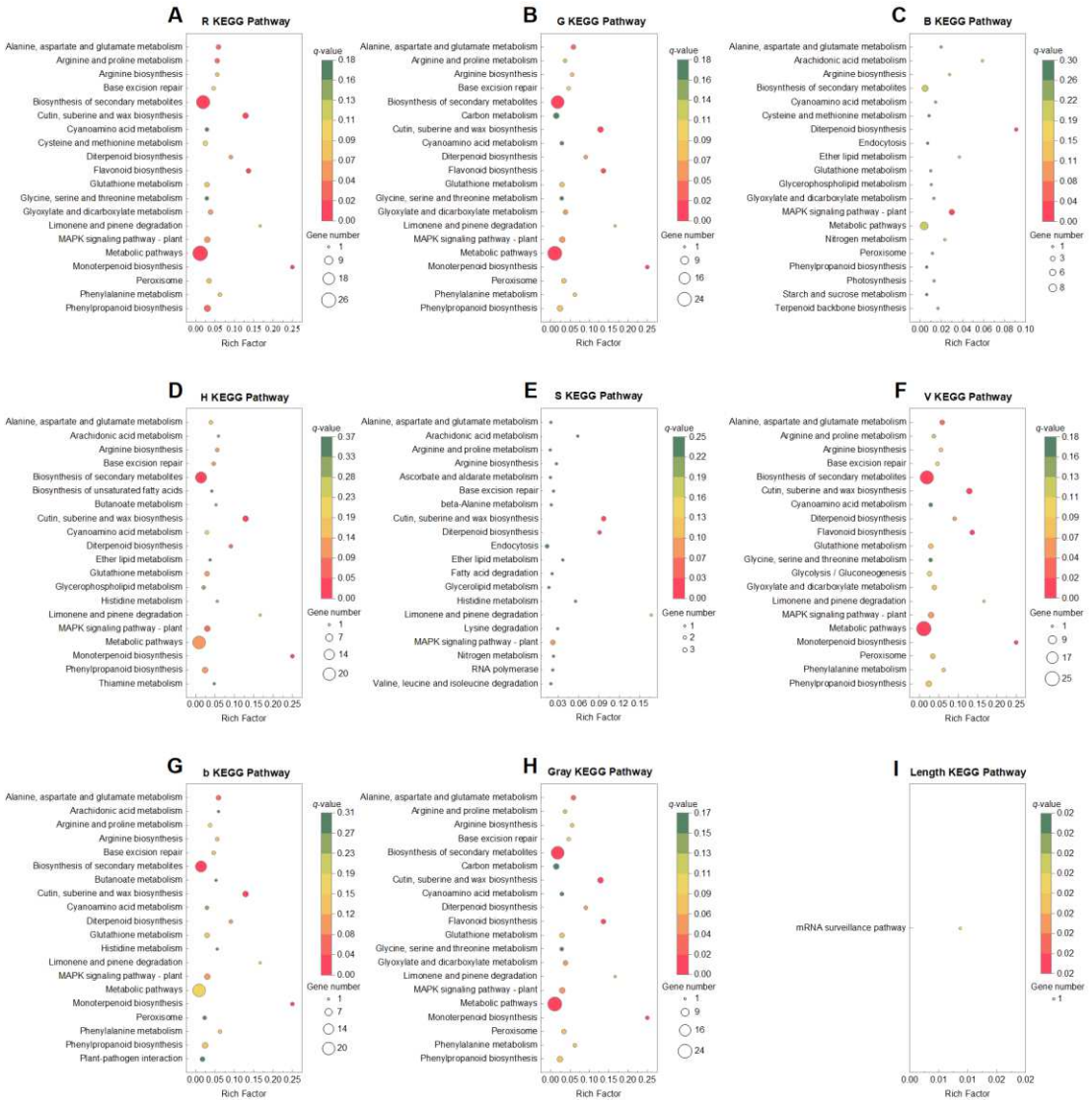

Figure 5 KEGG enrichment analysis of candidate genes for significant SNPs of color and morphological traits. A to I show the candidate gene enrichment in different pathways for R, G, B, H, S, V, b, Gray, and Length, respectively. Genes with $q < 0.05$ are significantly enriched, with p corrected using the Benjamini-Hochberg method.

DISCUSSION

In this study, we systematically analyzed the multidimensional phenotypic traits of a sheepgrass seed population by employing a high-throughput AIseed phenotyping platform in combination with SNP liquid chip genotyping. Through the integration of GWAS and KEGG enrichment analyses, we explored the genetic architecture and molecular mechanisms underlying variation in seed traits. The innovative application of the AIseed high-throughput phenotyping platform enabled the efficient collection of 54 key trait datasets across multiple dimensions—including seed morphology, color, and texture—from 262 accessions (Table 1). This approach overcomes the limitations

of traditional seed traits assessment, allowing for high-quality, cost-effective data collection using a flatbed scanner (Tu et al., 2023), and markedly enhances the accuracy and efficiency of trait extraction. Our results not only elucidate the genetic basis of phenotypic variation in sheepgrass seeds but also provide a theoretical framework and technical foundation for molecular breeding in forage grasses.

To address the prevalent issue of multicollinearity in high-dimensional phenotypic datasets (Figure 1), we adopted the elastic net regression method described by Cho et al. (2009) for trait selection (Figure 2). Compared to conventional linear regression approaches, the elastic net is more effective at identifying key traits. Our results indicate that seed coat texture and seed size—both reflecting seed plumpness—as well as seed coat greenness, which indicates maturity, are all significantly positively correlated with HGW. Additionally, seed coat color depth and maturity were found to be closely associated with GR. These findings are consistent with those reported by Gong et al. (2024) in peanut (*Arachis hypogaea* L.), yet differ from the negative correlations observed by Debeaujon et al. (2000) and Lepiniec et al. (2006) in *Transparent Testa* (*AtTT*) gene family mutants, suggesting that the genetic regulatory mechanisms controlling seed traits may differ across species.

The results of principal component analysis (Table 2 and Table 3) demonstrated that dehulling substantially increased the contribution of textural traits in germplasm evaluation (Table 4), thereby providing a basis for optimizing phenotypic testing protocols in forage seeds. Furthermore, the phenotypic evaluation system constructed in this study exhibited good applicability within the experimental population (Figure 3), offering a methodological reference for large-scale germplasm screening.

GWAS identified multiple significant SNP loci associated with color traits (Figure 4), and revealed genetic colocalization of different color traits at position 433,045,629 bp on chromosome Lc2Xm, suggesting that these traits may be regulated by a common gene network. KEGG enrichment analysis (Figure 5) further indicated that phenylpropanoid and flavonoid biosynthesis pathways are significantly enriched in the regulation of seed traits, consistent with previous studies on seed coat color diversity in soybean (*Glycine max* (L.) Merr.) (Fan et al., 2025) and common bean (*Phaseolus vulgaris* L.) (Roy et al., 2025). To our knowledge, this is the first study to demonstrate a direct association between these metabolic pathways and seed agronomic traits in sheepgrass.

During candidate gene mining, the F-box protein gene (*At5g07610*) was identified

as a significant association locus for multiple traits. This gene may participate in seed stress responses via the ubiquitination pathway (Malabarba et al., 2021) and regulate the transition between asexual and sexual reproduction in plants (Pablo Selva et al., 2020). This locus is located 772 bp upstream of the transcription start site and has a high minor allele frequency (MAF > 0.4), highlighting its potential for marker development in molecular breeding.

Five candidate genes in the phenylpropanoid biosynthesis pathway (*PAL*, *PER18*, *PER50*, *BGLU16*, and *BACOVA_02659*) encode key enzymes involved in secondary metabolism. *PAL* and *PER18/50* may influence seed coat coloration by regulating the metabolism of flavonoids, tannins, and reactive oxygen species, thereby controlling seed dormancy and germination (Ren et al., 2023; Li et al., 2025). *BGLU16* and *BACOVA_02659* may affect seed germination by modulating carbohydrate metabolism, including the conversion of polysaccharides to glucose (Wu et al., 2023; Li et al., 2025). In the flavonoid biosynthesis pathway, the discovery of the *ANR* gene cluster (Lc2Xm: 478.150 to 478.199 Mb) provides a molecular explanation for the positive correlation between seed coat color intensity and GR. Together with the findings of Gong et al. (2024) in peanut, this suggests that sheepgrass may modulate active oxygen metabolism through flavonoid biosynthesis, thereby influencing seed vigor. In future work, gene editing, transgenic approaches, and epigenetic modifications could be applied to enhance the activities of phenylpropanoid and flavonoid synthases and to increase the expression of stress-resistance genes, thereby improving seed stress tolerance and storage stability (Dong et al., 2021).

Moreover, the enrichment of genes involved in cutin, suberin, and wax biosynthesis supports the “endogenous cuticle” hypothesis proposed by De Giorgi et al. (2015), indicating that the physical barrier of the seed coat plays a crucial role in seed dormancy regulation. The *PXG* gene may indirectly affect seed traits by participating in the synthesis of cutin polyester precursors, thereby broadening our understanding of the genetic mechanisms underlying seed coat development.

In summary, the molecular-level findings of this study are consistent with the perspectives of Finkelstein et al. (2008) and Linkies et al. (2010), who highlighted the shared genetic regulatory mechanisms governing seed morphology, color, texture, dormancy, and germination. This work thus provides both phenotypic and genotypic evidence for the genetic improvement of sheepgrass germplasm. Nevertheless, the predictive accuracy of our GR model (PCA) requires further enhancement, and its

applicability to different ecotypes needs to be validated through multi-environment trials. Future research could draw on the approaches of Azam et al. (2024) and Morade et al. (2025) to integrate multidimensional data on forage yield, salt tolerance, and drought tolerance in sheepgrass, enabling the identification of elite germplasm with both high stress resistance and productivity. Additionally, the functional roles of several key candidate genes (e.g., *Rngtt* and *PXG*) remain unclear and warrant further investigation utilizing gene editing and related technologies. Collectively, these efforts will promote the transition from traditional, experience-based breeding to precision molecular design in sheepgrass, providing a robust scientific foundation for the sustainable development of grassland agriculture.

CONCLUSION

This study integrated high-throughput phenotypic data generated by the AIseed platform with genomic data obtained from a 50K SNP array to elucidate the genetic associations between seed morphological, color, and textural traits, and key agronomic traits such as HGW and GR in sheepgrass. By applying elastic net regression, we identified key phenotypic indicators associated with these agronomic traits. PCA achieved a 71% accuracy rate in high germination rate germplasm screening, demonstrating the effectiveness of the phenotypic evaluation system. GWAS mapped significant SNP loci for color and morphological traits primarily to the Lc2Xm and Lc6Xm chromosomes and highlighted the central regulatory roles of the phenylpropanoid and flavonoid biosynthesis pathways—including candidate genes such as *PAL* and *ANR*. These findings establish an efficient phenotypic screening platform and provide valuable genetic resources for molecular breeding in sheepgrass, offering significant potential for the genetic improvement of forage grasses.

FUNDING

This work was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences, grant number XDA26030202, and the National Key Research and Development Program, grant number 2022YFF1003203.

ACKNOWLEDGMENTS

The authors declare no conflicts of interest and affirm their agreement with the content of this manuscript.

AUTHOR CONTRIBUTIONS

S.C. and G.L. conceived the research and designed the methodology. Q.S provided technical support for seed image-traits scanning. L.T., C.K., and S.H. conducted experiments and generated phenotypic data. L.T., and C.K. performed data analysis and visualization. L.T., S.H., X.H., and L.C collected and prepared experimental materials. L.T. and S.C. developed the graphs and drafted the manuscript.

REFERENCE

Allen A M, Winfield M O, Burridge A J, et al. Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*)[J]. Plant Biotechnology Journal, 2017, 15(3): 390-401.

Azam M G, Sarker U, Hossain M A, et al. Phenotypic diversity in qualitative and quantitative traits for selection of high yield potential field pea genotypes[J]. Scientific Reports, 2024, 14(1): 18561.

Blanchy G, Deroo W, De Swaef T, et al. Closing the phenotyping gap with non-invasive belowground field phenotyping[J]. Soil, 2025, 11(1): 67-84.

Bu D, Luo H, Huo P, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis[J]. Nucleic Acids Research, 2021, 49(W1): W317-W325.

Burridge A J, Winfield M, Przewieslik-Allen A, et al. Development of a next generation SNP genotyping array for wheat[J]. Plant Biotechnology Journal, 2024, 22(8): 2235-2247.

Chen S, Huang X, Yan X, et al. Transcriptome Analysis in Sheepgrass (*Leymus chinensis*): A Dominant Perennial Grass of the Eurasian Steppe[J]. PLoS One, 2013, 8(7): e67974.

Chen S, Jia J, Cheng L, et al. Transcriptomic Analysis Reveals a Comprehensive Calcium- and Phytohormone-Dominated Signaling Response in *Leymus chinensis* Self-Incompatibility[J]. International Journal of Molecular Sciences, 2019, 20(9): 2356.

Cho S, Kim H, Oh S, et al. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis.[J]. BMC proceedings, 2009, 3 Suppl 7: S25.

De Giorgi J, Piskurewicz U, Loubery S, et al. An Endosperm-Associated Cuticle Is Required for Arabidopsis Seed Viability, Dormancy and Early Control of Germination[J]. PLoS Genetics, 2015, 11(12): e1005708.

Debeaujon I, Léon-Kloosterziel K M, Koornneef M. Influence of the testa on seed dormancy, germination, and longevity in *Arabidopsis*[J]. Plant Physiology, 2000, 122(2): 403-413.

Dong N Q, Lin H X. Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions[J]. Journal of Integrative Plant Biology, 2021, 63(1): 180-209.

Doyle J J, Doyle J L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue[J]. Phytochemical Bulletin, 1987, 19: 11-15.

Fan Y, Hussain S, Wang X, et al. Metabolomic and Transcriptomic Analyses of Flavonoid Biosynthesis in Different Colors of Soybean Seed Coats[J]. International Journal of Molecular Sciences, 2025, 26(1): 294.

592 Finkelstein R, Reeves W, Ariizumi T, et al. Molecular aspects of seed dormancy[J]. Annual Review
593 of Plant Biology, 2008, 59: 387-415.
594 Gong F, Cao D, Li Zhuo, et al. Metabolomic and Transcriptomic Analysis Reveals Flavonoid-
595 Mediated Regulation of Seed Antioxidant Properties in Peanut Seed Vigor[J]. Antioxidants,
596 2024, 13(12): 1497.
597 Guo Z, Yang Q, Huang F, et al. Development of high-resolution multiple-SNP arrays for genetic
598 analyses and molecular breeding through genotyping by target sequencing and liquid chip[J].
599 Plant Communications, 2021, 2(6): 100230.
600 Hu Y, Liu Y, Wei J J, et al. Regulation of seed traits in soybean.[J]. aBIOTECH, 2023, 4(4): 372-
601 385.
602 Krzyszton M, Sacharowski S P, Manjunath V H, et al. Dormancy heterogeneity among *Arabidopsis*
603 *thaliana* seeds is linked to individual seed size[J]. Plant Communications, 2024, 5(2): 100732.
604 Lepiniec L, Debeaujon I, Routaboul J M, et al. Genetics and biochemistry of seed flavonoids[J].
605 Annual Review of Plant Biology, 2006, 57: 405-430.
606 Li T, Tang S, Li W, et al. Genome evolution and initial breeding of the Triticeae grass *Leymus*
607 *chinensis* dominating the Eurasian Steppe[J]. Proceedings of the National Academy of
608 Sciences, 2023, 120(44): e2308984120.
609 Li Y, Li X, Li F, et al. Comparative transcriptome analysis reveal gene regulation of dormancy
610 release in Cardiocrinum giganteum seeds induced by temperature[J]. Frontiers in Plant Science,
611 2025, 16: 1-19.
612 Linkies A, Graeber K, Knight C, et al. The evolution of seeds[J]. New Phytologist, 2010, 186(4):
613 817-831.
614 Liu Y, Liu S, Zhang Z, et al. GenoBaits Soy40K: a highly flexible and low-cost SNP array for
615 soybean studies[J]. Science China Life Sciences, 2022, 65(9): 1898-1901.
616 Liu Z, Chen Y, Meng J, et al. Seed shattering and relevant traits of *Leymus chinensis*[J]. Acta
617 Agrestia Sinica, 2013, 21(1): 152-158
618 Ma J, Cao Y, Wang Y, et al. Development of the maize 5.5K loci panel for genomic prediction
619 through genotyping by target sequencing[J]. Frontiers in Plant Science, 2022, 13: 972791.
620 Mahajan S, Das A, Sardana H K. Image acquisition techniques for assessment of legume quality[J].
621 Trends in Food Science & Technology, 2015, 42(2): 116-133.
622 Mahajan S, Mittal S K, Das A. Machine vision based alternative testing approach for physical purity,
623 viability and vigour testing of soybean seeds (Glycine max)[J]. Journal of Food Science and
624 Technology-Mysore, 2018, 55(10): 3949-3959.
625 Malabarba J, Windels D, Xu W, et al. Regulation of DNA (de)Methylation Positively Impacts Seed
626 Germination during Seed Development under Heat Stress[J]. Genes, 2021, 12(3): 457.
627 McDonald M B. Seed quality assessment[J]. Seed Science Research, 1998, 8(2): 265-275.
628 Morade A S, Sharma R M, Dubey A K, et al. Phenotyping drought stress tolerance in citrus
629 rootstocks using high-throughput imaging and physio-biochemical techniques[J]. BMC Plant
630 Biology, 2025, 25(1): 753.
631 Nonogaki H. Seed germination and dormancy: The classic story, new puzzles, and evolution[J].
632 Journal of Integrative Plant Biology, 2019, 61(5): 541-563.
633 Oh M A, Park J E, Kim J Y, et al. Seed traits inheritance in Fagopyrum esculentum Moench. based

on image analysis method[J]. Frontiers in Plant Science, 2024, 15: 1445348.
Pablo Selva J, Zappacosta D, Carballo J, et al. Genes Modulating the Increase in Sexuality in the Facultative Diplosporous GrassEragrostis curvulaunder Water Stress Conditions[J]. Genes, 2020, 11(9): 969.
Qi J, Gao F, Wang Y, et al. Multiscale phenotyping of grain crops based on three-dimensional models: A comprehensive review of trait detection[J]. Computers and Electronics in Agriculture, 2025, 237: 110597.
Ren Y, Shen F, Liu J, et al. Application of Methionine Increases the Germination Rate of Maize Seeds by Triggering Multiple Phenylpropanoid Biosynthetic Genes at Transcript Levels[J]. Plants-Basel, 2023, 12(22): 3802.
Roth L, Marzougui A, Walter A. A review of the journey of field crop phenotyping: From trait stamp collections and fancy robots to phenomics-informed crop performance predictions[J]. Journal of Plant Physiology, 2025, 311: 154542.
Roy J, Sreedasyam A, Osborne C, et al. Seed coat transcriptomic profiling of 5-593, a genotype important for genetic studies of seed coat color and patterning in common bean (*Phaseolus vulgaris* L.)[J]. BMC Plant Biology, 2025, 25(1): 284.
Tang Z, Chen Z, Gao Y, et al. A Strategy for the Acquisition and Analysis of Image-Based Phenome in Rice during the Whole Growth Period[J]. Plant Phenomics, 2023, 5: 0058.
Tu K, Wu W, Cheng Y, et al. AIseed: An automated image analysis software for high-throughput phenotyping and quality non-destructive testing of individual plant seeds[J]. Computers and Electronics in Agriculture, 2023, 207: 107740.
Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data[J]. Nucleic Acids Research, 2010, 38(16): e164.
Watt M, Fiorani F, Usadel B, et al. Phenotyping: New Windows into the Plant for Breeders[J]. Annual Review of Plant Biology, 2020, 71: 689-712.
Wu T, Dai J, Shen P, et al. Seedscreener: A novel integrated wheat germplasm phenotyping platform based on NIR-feature detection and 3D-reconstruction[J]. Computers and Electronics in Agriculture, 2023, 215: 108378.
Wu X, He Q, Wang Z, et al. Differential expression analysis of genes related to starch and sucrose metabolic pathways in potato tubers under drought stress simulated based on transcriptome sequencing[J]. Genomics and Applied Biology, 2023, 42(1): 44-59.
Wu Z, Hou X, Ren W, et al. Prediction of the potential geographic distribution of *Leymus chinensis* based on MaxEnt and collection and protection of germplasm[J]. Acta Prataculturae Sinica, 2018, 27(10): 125-135.
Xu S, Huang X, Liang X, et al. Application of wearable sensors in crop phenotyping and microenvironment monitoring[J]. Chemical Engineering Journal, 2025, 505: 159059.
Yang W, Liu S, Yuan G, et al. Germination characteristics among different sheepgrass (*Leymus chinensis*) germplasm during the seed development and after-ripening stages[J]. PeerJ, 2019, 7: e6688.
Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies[J]. Nature Genetics, 2012, 44(7): 821-U136.