GWAS-informed genomic selection for cold tolerance in pepper (Capsicum annuum L.)

Kyeongseok Lee
Seoul National University
Geon Woo Kim
Seoul National University
Hee-Jin Jeong
Nongwoo Seed Turkey
Hyeon-Seok Jeong
Nongwoo Bio
Jin-Kyung Kwon
Seoul National University
Byoung-Cheorl Kang
bk54@snu.ac.kr

Seoul National University

Research Article

Keywords: Pepper, Cold tolerance, Cold-tolerance index (CTI), Genome-wide association study (GWAS), Genomic selection (GS)

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9143989/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Plant Biology on July 7th, 2026. See the published version at https://doi.org/10.1186/s12870-026-09438-1.

GWAS-informed genomic selection for cold tolerance in pepper (*Capsicum annuum* L.)

Kyeongseok Lee[1], Geon Woo Kim[1], Hee-Jin Jeong[2], Hyeon-Seok Jeong[3], Jin-Kyung Kwon[1], and Byoung-Cheorl Kang[1*]

[1] Department of Agriculture, Forestry and Bioresources, Research Institute of Agriculture and Life Sciences, Plant Genomics and Breeding Institute, College of Agriculture and Life Sciences, Seoul National University, Seoul, 08826, South Korea

[2] Breeding Research Center, Nongwoo Seed Turkey, Muratpaşa, 07230, Türkiye

[3] Breeding Research Center, Nongwoo Bio, Yeoju, 12655, South Korea

[*] Corresponding author: Byoung-Cheorl Kang

Tel: +82-2-880-4563; Fax: +82-2-873-2056; E-mail: bk54@snu.ac.kr; Orcid: 0000-0002-7606-7258

Abstract

Background

Breeding for cold tolerance in pepper (*Capsicum annuum* L.) is critical to mitigate yield losses caused by unpredictable temperature fluctuations associated with climate change. However, genetic improvement of this trait is hindered by challenges in accurate phenotyping, particularly at the adult stage, and by its complex genetic architecture involving numerous minor-effect loci. While genomic selection (GS) offers a promising solution to accelerate genetic gain, its predictive ability is often limited by statistical noise from uninformative markers within whole-genome marker sets. This study aimed to overcome this limitation by developing a robust phenotypic index and implementing a genome-wide association study (GWAS)-informed GS strategy.

Results

We phenotyped 192 pepper accessions from a core collection for cold tolerance using a visual survival score (Surv) and a newly developed composite cold-tolerance index (CTI). Both CTI (h^2 = 0.55) and

Surv (h^2 = 0.53) showed moderate heritability, suggesting a substantial contribution from additive genetic variance to the phenotypic variation of cold tolerance in adult plants. GWAS identified 13 candidate genomic regions associated with cold tolerance; these regions included *TRM9*, *CAP1*, and *PP2A-2*, genes previously implicated in abiotic stress responses. A GWAS-informed GS model using a selected subset of 1,024 markers achieved a prediction accuracy of 0.78, representing a substantial improvement over that obtained with the standard model using the full marker set of 73,502 markers (0.203). Notably, a control model using a random marker set of identical size (1,024 markers) yielded an accuracy of only 0.122, confirming that the greater predictive power of the GWAS-informed GS model was driven by the genetic relevance of the selected markers rather than by lower marker density.

Conclusions

Our study demonstrates that assessing cold tolerance via the CTI is pivotal for overcoming the limitations of small-scale sample collection. By turning ordinal data into a continuous spectrum, the CTI can effectively unmask hidden genetic variation. Integrating this refined phenotype with GWAS-informed marker selection into prediction models significantly enhanced the accuracy of genomic prediction for cold tolerance in adult pepper plants. This integrated framework offers a practical and efficient roadmap for accelerating breeding cycles and improving selection precision for complex abiotic stress traits in pepper breeding.

Keywords: Pepper, Cold tolerance, Cold-tolerance index (CTI), Genome-wide association study (GWAS), Genomic selection (GS)

Background

Pepper (*Capsicum annuum* L.) is a crop of significant global economic value, with annual production reaching over 44 million tons in 2023 (FAOSTAT). However, this production is increasingly threatened by climate change, which has intensified the frequency of unpredictable weather events [1]. While global warming is a primary concern, the associated climatic volatility also triggers sudden temperature

53 drops, posing severe risks to crop stability [2, 3]. Exposure to cold stress impairs photosynthesis and
54 disrupts metabolism, necessitating the development of varieties with enhanced cold tolerance [4].

55 Breeding for greater tolerance to abiotic stresses is challenging due to the complex genetic
56 architecture of these traits. Cold tolerance is polygenic, controlled by the cumulative action of numerous
57 minor-effect loci and often influenced by interactions with the environment [5]. Over the past decade,
58 significant progress has been made in dissecting the genetic basis of abiotic stress tolerance in major
59 model plants and staple crops. In rice (*Oryza sativa*), the cloning of *COLD1* revealed a critical calcium
60 (Ca^{2+}) signaling mechanism at the plasma membrane that triggers cold defense responses [6]. Within
61 the Solanaceae, research in tomato (*Solanum lycopersicum*) identified the B-box gene *SlBBX31* as a
62 master regulator that integrates light and temperature signals to enhance cold tolerance [7]. These
63 studies emphasize how mapping the loci underlying phenotypic variation can successfully resolve the
64 polygenic nature of cold tolerance, providing a foundation for more targeted breeding approaches such
65 as marker-assisted selection (MAS).

66 MAS has yielded success in specific instances, including for the *Na^+ exclusion 2* (*Nax2*) locus
67 for sodium (Na^+) exclusion in wheat (*Triticum aestivum*) [8] or the *zmm28* regulator which enhances
68 grain yield through improved photosynthesis and nitrogen utilization in maize (*Zea mays*) [9]. However,
69 using such locus-specific strategies to breed for stress tolerance is often insufficient for traits where
70 variation is distributed across the entire genome. Indeed, in most cases, the few known quantitative trait
71 loci (QTLs) explain only a small fraction of the total phenotypic variance, raising the problem of
72 missing heritability [10]. Consequently, effective improvement of complex traits like cold tolerance
73 requires a shift from selecting individual loci to integrating effects across the genome.

74 To address the limitations of MAS for complex traits, genomic selection (GS) was introduced to
75 predict genomic estimated breeding values (GEBVs), which account for both major and minor effects
76 across the genome [13]. GS has proven powerful for improving complex traits across a wide range of
77 crops such as wheat, maize, soybean (*Glycine max*), and pepper [14-17]. In pepper, foundational studies
78 have established the practicability of GS for agricultural traits. Hong et al. [17] applied GS to fruit-

related traits by training prediction models on a core collection with genome-wide SNP markers [18], evaluating performance via cross-validation. They further examined how marker density and population structure influence prediction accuracy and validated the trained models in an independent recombinant inbred line population. In the case of cold stress, GS has been applied to wheat at the seedling stage to predict cold tolerance [20], and genetic loci associated with cold stress have been identified for maize at the seedling stage and in tomato at the adult stage [26, 27].

Despite these successes, the prediction accuracy of standard GS approaches is limited by the inclusion of many polymorphic sites (or markers) that are unassociated with the trait of interest, thus introducing statistical noise. A genome-wide association study (GWAS) can be used to address this issue by serving as a statistical filter to identify the most relevant markers, in addition to its conventional role in candidate gene identification [19]. This GWAS-informed approach to GS can yield higher prediction accuracy or achieve comparable accuracy with greater efficiency by focusing on pre-selected subsets of trait-associated markers [20-23]. For example, Kim et al. [21] improved genomic prediction for capsaicinoid content in pepper by incorporating markers derived from GWAS either as a reduced marker set for model training or as fixed effects within GS models, thereby significantly enhancing prediction accuracy.

To fully leverage the power of GS and GWAS, the phenotypic data used must be as robust as the genotypic data. Quantifying a complex biological response like cold tolerance often relies on discrete, ordinal scales, such as visual assessments. While practical, such methods cannot fully capture the continuous spectrum of a plant's status. Furthermore, this type of data does not follow a normal distribution, which can limit the statistical power and precision of downstream quantitative genetic analysis that is necessary to dissect polygenic traits [11, 12]. To overcome the limitations of discrete scores, an effective strategy is to develop a composite index. The utility of such composite indices for enhancing the genetic analysis of complex traits has been demonstrated for rice, maize, and wheat in previous studies [24, 25]. This approach involves integrating a key indicator of the trait of interest with several other supporting traits, which are typically selected based on their strong statistical relationship

with the primary trait, such as a high and significant Pearson's correlation coefficient. The goal of generating an index is to convert the raw phenotypic data into a more continuous and normally distributed set of values that is better suited for the statistical assumptions of GS prediction models, thereby enhancing their potential accuracy.

In pepper, research on the genetic basis of cold tolerance has focused on transcriptomic analysis to elucidate cold stress responses at the seedling stage [28]. A GWAS-informed GS strategy has yet to be implemented for enhancing cold tolerance in pepper specifically at the adult stage, a phase directly critical for fruit quality and recovery. The aims of this study were to: (1) develop and validate a composite cold tolerance index (CTI) to overcome phenotyping challenges; (2) use this CTI to identify significant genomic regions and candidate genes associated with cold tolerance using GWAS; and (3) systematically evaluate whether a GWAS-informed marker selection strategy can optimize GS prediction accuracy over a strategy based on the full marker set. This study is expected to provide a more efficient and precise molecular breeding strategy for the development of cold-tolerant pepper cultivars.

Results

Phenotypic evaluation and year-to-year trait stability

We collected phenotypic data for 14 traits over two consecutive growing seasons. The survival score (Surv) for cold tolerance across the 192 *C. annuum* accessions predominantly clustered between values of 3 and 5, reflecting a moderate tolerance to cold stress (Fig. 1A). While 185 accessions from the core collection showed growth cessation and fruit drop under cold stress, seven accessions displayed superior recovery from cold stress, with Surv scores of 7. Notably, no accession within the core collection reached the maximum possible Surv value of 9; the highly cold-tolerant cultivar 'Pusula' did, however, serving as a control for full recovery.

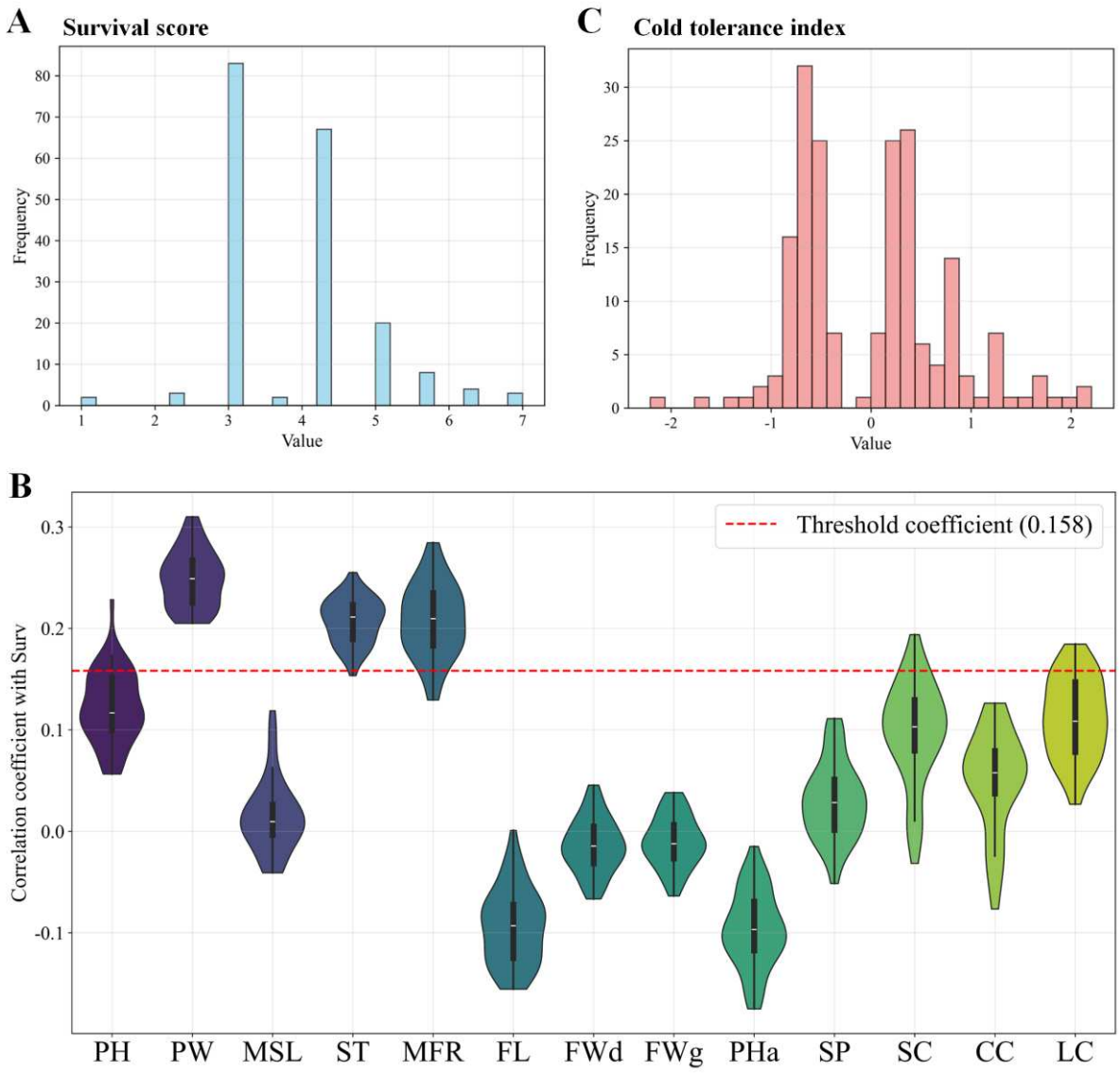

Figure 1. Phenotypic overview and derivation of the cold-tolerance index. (A) Distribution of values of survival score (Surv). (B) Violin plot showing the distribution of cross-validated Pearson's correlation coefficients between Surv and each of the 13 traits measured. The width of each violin indicates the density of the data, and the central line represents the median correlation coefficient. The dashed red line represents the critical significance threshold (p = 0.05). PH, plant height; PW, plant width; MSL, main stem length; ST, stem thickness; MFR, mature fruit ratio; FL, fruit length; FWd, fruit width; FWg, fruit weight; PHa, plant habit; SP, stem pubescence; SC, stem color; CC, cotyledon color; LC, leaf color. (C) Distribution of the calculated values for the cold-tolerance index (CTI).

We assessed the stability of the above responses by calculating the Pearson's correlation coefficients for each trait between the 2023–2024 and 2024–2025 growing seasons, yielding coefficients ranging from $r = 0.13$ to $r = 0.99$ (Fig. S1). Mature fruit ratio (MFR) and cotyledon color (CC) were excluded from this year-to-year analysis, as they were only evaluated during one growing season. We observed high phenotypic stability for leaf color (LC, $r = 0.99$), fruit width (FWd, $r = 0.97$), plant habit (PHa, $r = 0.97$), fruit weight (FWg, $r = 0.96$), fruit length (FL, $r = 0.92$), Surv ($r = 0.89$), plant height (PH, $r = 0.86$), and main stem length (MSL, $r = 0.79$). Certain vegetative traits showed less reproducible phenotypic values from one year to the next, such as stem pubescence (SP, $r = 0.57$), plant width (PW, $r = 0.46$), stem thickness (ST, $r = 0.38$), and stem color (SC, $r = 0.13$), indicating that these raw phenotypic measurements from individual years may be confounded by environmental effects, necessitating a statistical framework to minimize year-specific bias and accurately estimate the genetic mean. To address this and obtain representative phenotypic values that accurately reflect genetic potential, we estimated the best linear unbiased predictions (BLUPs) for continuous agronomic traits and calculated the arithmetic means for categorical traits across the two seasons. This statistical adjustment corrected for variation caused by year and replication, ensuring reliable data from even environmentally sensitive traits for genetic analysis.

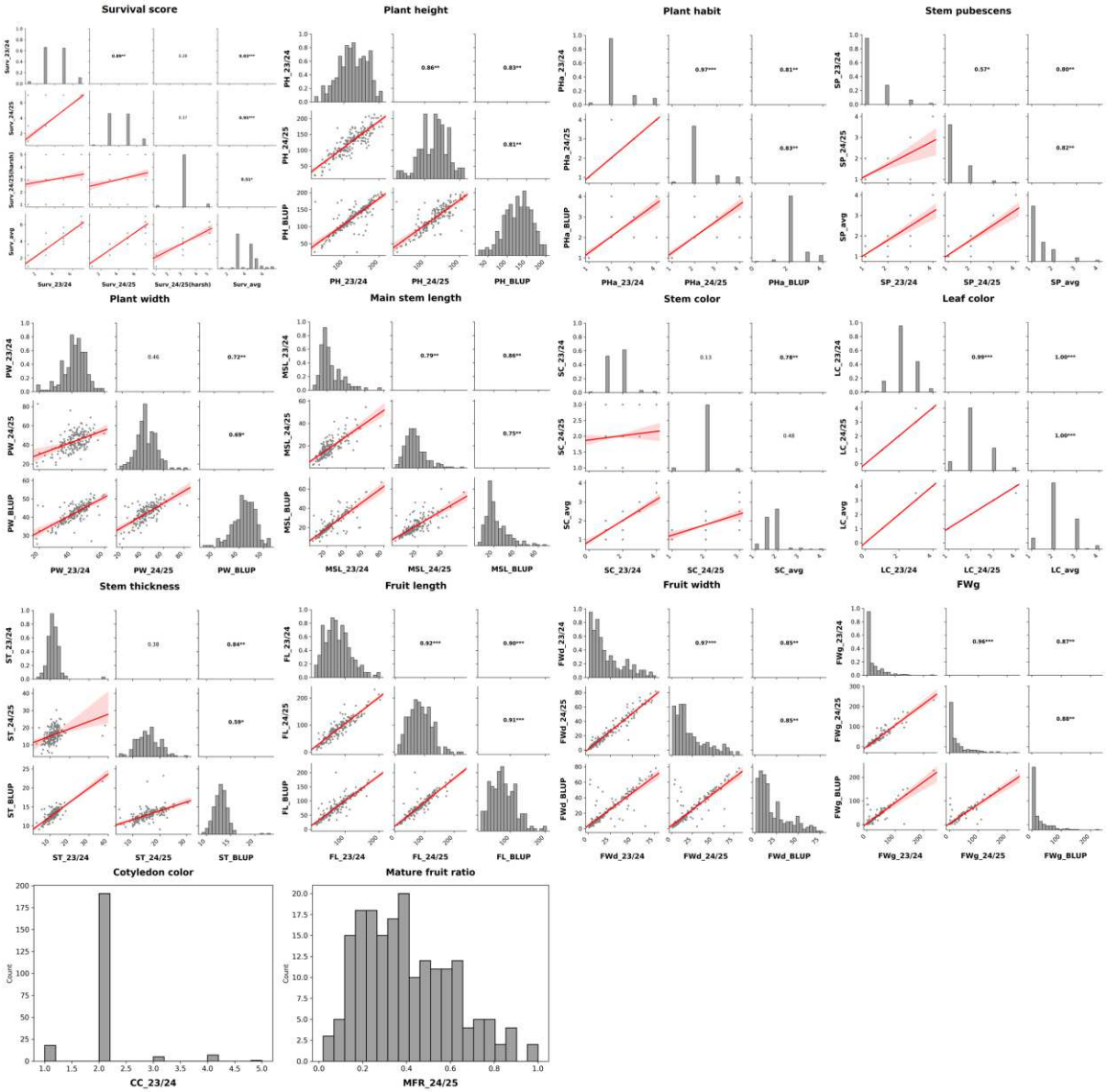

Fig. S1. Year-to-year correlations and distributions of traits evaluated across two growing seasons. This figure shows the reproducibility and stability of phenotypic measurements between years (2023/24 and 2024/25) for each trait, as well as the correlation with BLUP values. For each trait, histograms on the diagonal show the distributions of values from the 2023/24 season, the 2024/25 season, and the BLUP values, while the other panels show scatter plots between two variables with fitted linear regression lines. Pearson correlation coefficients (*r*) are shown in each scatter plot panel, and asterisks indicate the statistical significance of the correlation (*; $p < 0.05$, **; $p < 0.01$, ***; $p < 0.001$). Cotyledon color (CC) and mature fruit ratio (MFR) were evaluated over a single growing season.

Development of the cold-tolerance index via correlation-based trait selection

To overcome the limitations of the discrete Surv scale and capture physiological stress responses as a continuous spectrum, we developed a composite cold-tolerance index (CTI). A prerequisite for this development was the identification of robust traits that consistently reflect plant cold tolerance. Using a rigorous selection procedure involving 50 iterations of five-fold cross-validation, we evaluated each of the 13 agronomic traits for their potential correlation with Surv. Among the traits, PW, ST, and MFR were the most reliable indicators of Surv, maintaining Pearson's correlation coefficients with Surv above the significance threshold ($p < 0.05$) in 100%, 100%, and 96% of the validation folds, respectively (Table 1 and Fig. 1A). We excluded the remaining 10 traits, as they failed to reach the 80% threshold.

Table 1. Pearson's correlation coefficient–based trait selection for CTI processing.

Trait	Mean	SD	Significance ($p < 0.05$)	
			Count	Ratio
PH	0.1222	0.0354	18	0.36
PW	0.2486	0.0279	50	1
MSL	0.0146	0.0352	0	0
ST	0.2074	0.0223	50	1
MFR	0.2079	0.0362	48	0.96
FL	−0.0936	0.0369	0	0
FWd	−0.0140	0.0266	0	0
FWg	−0.0102	0.0251	0	0
PHa	−0.0970	0.0373	0	0
SP	0.0299	0.0377	0	0
SC	0.0978	0.0499	8	0.16
CC	0.0504	0.0496	0	0
LC	0.1116	0.0393	14	0.28

PH, plant height; PW, plant width; MSL, main stem length; ST, stem thickness; MFR, mature fruit ratio; FL, fruit length; FWd, fruit width; FWg, fruit weight; PHa, plant habit; SP, stem pubescence; SC, stem color; CC, cotyledon color; LC, leaf color; SD, standard deviation; CTI, cold-tolerance index.

Note: Only traits with significance ratio > 0.8 were selected for CTI calculation.

Based on these results, we calculated the CTI as a weighted sum where Surv was assigned a dominant weight ($\alpha = 0.706$), while the supporting traits PW (0.073), ST (0.061), and MFR (0.061) contributed to the remaining proportion. This integration of PW, ST, and MFR together with Surv

189 modified the distribution of cold tolerance across the population (Fig. 1B, C). While Surv exhibited a
190 skewed and truncated distribution, with accessions primarily displaying coarse ordinal scores from 3 to
191 5, the CTI followed a more continuous distribution approaching normality. The CTI may thus
192 effectively unmask hidden phenotypic variance that was previously indistinguishable within the discrete
193 visual categories of Surv, providing a more refined phenotypic basis for quantitative genetic analysis.
194
195 **Estimates of heritability and distribution of polygenic variance**
196 To dissect the genetic basis of cold tolerance at the adult stage, we estimated single nucleotide
197 polymorphism (SNP)-based heritability (h^2) and partitioned the genetic variance explained across each
198 of the 12 chromosomes. We also tested the effect of allele frequency on phenotypic variance explained.
199 We determined that the cold-tolerance trait has a moderate heritability, with additive genetic variance
200 accounting for a substantial proportion of the observed phenotypic variation. The CTI yielded a
201 heritability estimate of 0.55 (± standard error [SE] of 0.17), which was comparable to that calculated
202 with discrete Surv ($h^2 = 0.53 \pm 0.18$). These results indicate that converting a discrete ordinal score into
203 a continuous CTI preserves the heritability of cold tolerance, demonstrating that the CTI was
204 successfully converted from discrete Surv retaining the genetic characteristics of Surv.
205 When we partitioned the variance across the 12 chromosomes of the pepper genome, we observed
206 that the genetic architecture of cold tolerance is characterized by widespread contributions across the
207 genome, supporting the polygenic nature of this trait (Fig. 2A). Chromosomes 3 and 4 were the greatest
208 contributors to the explained variance for Surv and the CTI, collectively explaining a larger fraction of
209 the total genetic variance than the other chromosomes. We then partitioned SNPs significantly
210 associated with cold tolerance as a function of allele frequency. We categorized SNPs into three bins:
211 low frequency (0.05–0.10), intermediate frequency (0.10–0.30), and high frequency (0.30) (Fig. 2B).
212 Much of the heritability was driven by SNPs with low- to intermediate-frequencies, which collectively
213 contributed to a heritability of 0.52. By contrast, high-frequency SNPs accounted for a smaller
214 proportion of the variance ($h^2 = 0.37$). This distribution pattern, where numerous variants with lower

frequencies explain the bulk of the genetic variance, underscores the complex and highly polygenic architecture of cold tolerance in pepper.

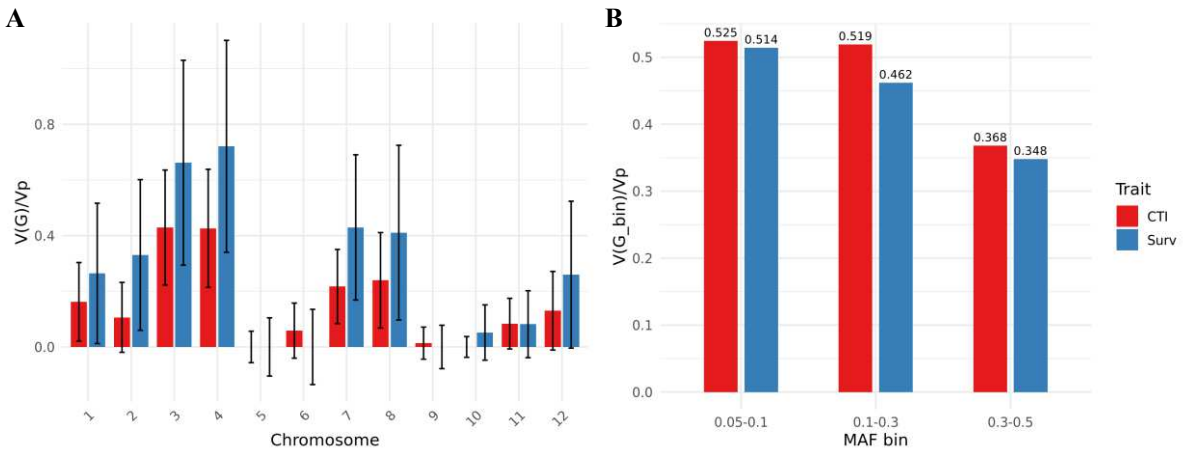

Figure 2. Chromosome-level and contributions of different MAF bins to variance for CTI and Surv traits based on GCTA GREML single-GRM analysis. Results related to the CTI are shown in red; results related to Surv are shown in blue. (A) Proportion of genetic variance explained by markers on each of the 12 chromosomes of the pepper genome. Data are presented as means ± standard errors. (B) Proportion of genetic variance relative to the total phenotypic variance for three minor allele frequency (MAF) bins. The numbers above the bars indicate the estimated proportion of variance explained for each bin size. The values are shown as proportions ranging from 0 to 1.

Genome-wide association signals and identification of candidate genes

We performed GWAS to explore the genomic regions associated with cold tolerance and identify an informative marker set for subsequent implementation in GS models. As phenotypes, we focused on Surv and the CTI. We applied the BLINK model, which identified significant associations that were undetectable by the six other models FarmCPU, MLMM, MLM, CMLM, GLM, and SUPER (Figs. 3, S3). A total of 13 unique SNPs surpassed the Bonferroni-corrected significance threshold ($p < 6.803 \times 10^{-7}$; Table 2). The Manhattan plots for Surv and the CTI showed peaks on chromosomes 4, 7, and 11 (Fig. 3).

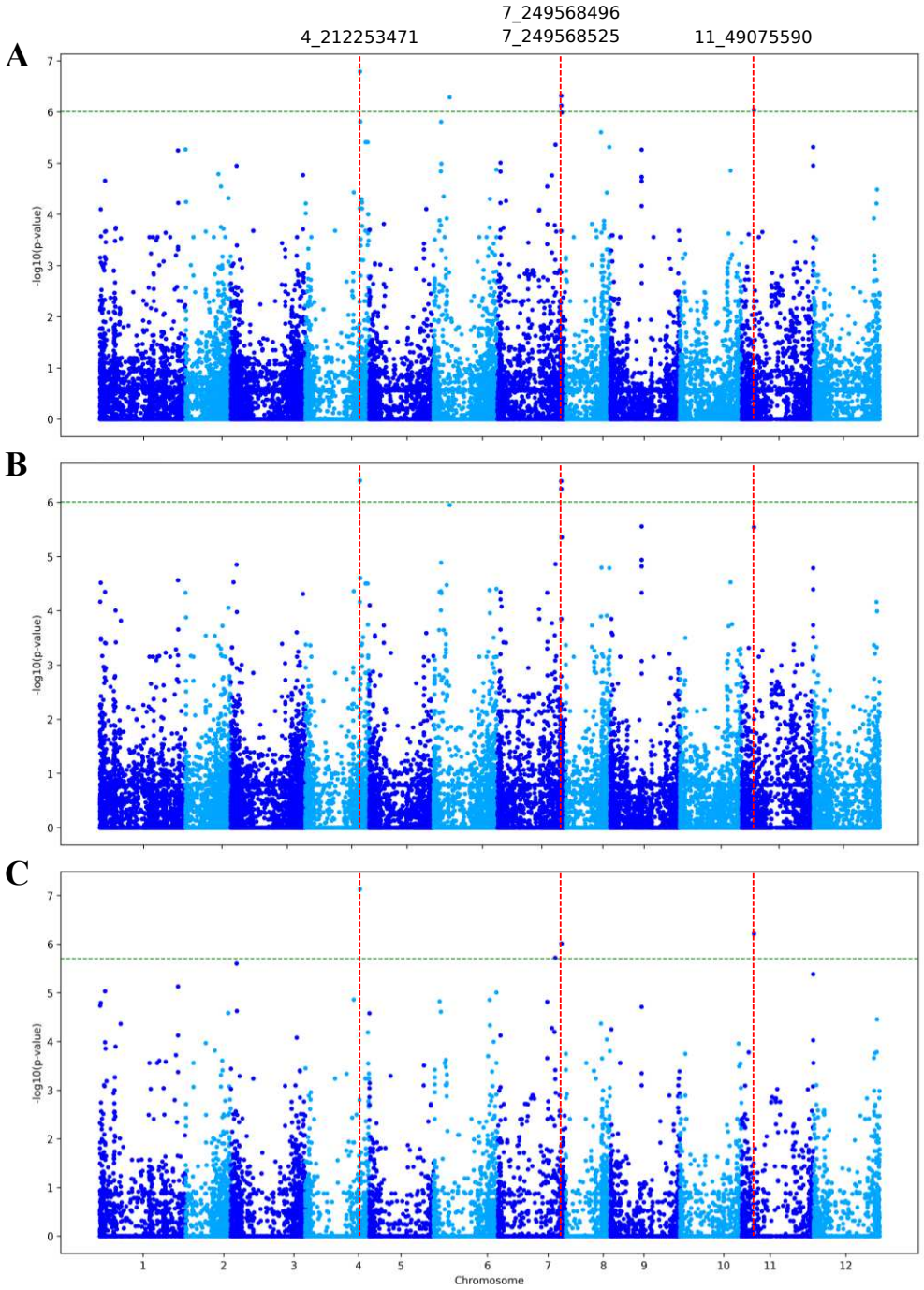

Figure 3. GWAS signals for cold tolerance. (A–C) Manhattan plots from GWAS for Surv and CTI, applying different SNP call-rate filters: (A) Surv (≥ 70%), (B) CTI (≥ 70%), and (C) CTI (≥ 90%). The red dashed vertical lines indicate genomic regions above the Bonferroni threshold (green dashed horizontal lines) on chromosomes 4, 7, and 11, detected for both Surv and CTI.

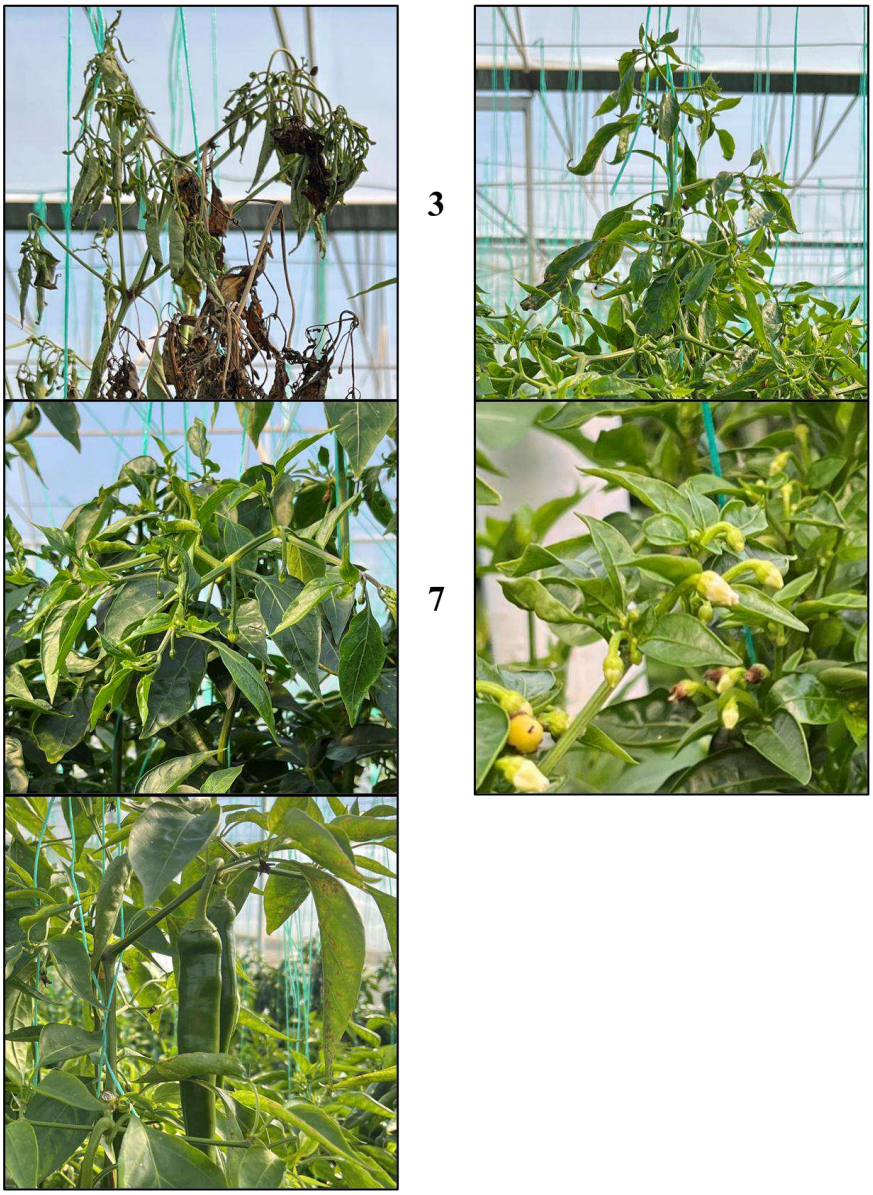

Fig. S3. Nine-point visual scoring scale for estimating the cold survival score. Representative images show the ordinal scale used for Surv evaluation: 1, plant death; 3, growth stops or becomes very slow, with flower and fruit drop; 5, growth is maintained, but flower and fruit drop occurs; 7, growth continues without flower or fruit drop; 9, growth is better than average under cold stress.

Table 2. Significant SNPs for survival score and CTI from GWAS using genotypic data based on three different SNP call-rate thresholds.

Trait	Chr.	Pos. (bp)	Call rate	Model	*P*-value	MAF	Effect
Surv	1	304,295,562	0.9	BLINK	1.40E−06	0.021	0.98
		304,295,609	0.9	BLINK	1.40E−06	0.021	0.98
	4	212,253,471	0.5	BLINK	5.07E−07	0.099	0.58
		212,253,471	0.7	BLINK	1.62E−07	0.099	0.57
		212,253,471	0.9	BLINK	2.96E−08	0.099	0.57
	6	63,882,822	0.7	BLINK	5.16E−07	0.013	1.44
	7	226,350,056	0.9	BLINK	7.92E−07	0.206	−0.42
		249,568,496	0.7	BLINK	7.43E−07	0.339	−0.34
		249,568,525	0.7	BLINK	4.80E−07	0.336	−0.34
		250,892,132	0.9	BLINK	2.02E−07	0.018	1.18
		250,896,197	0.9	BLINK	2.02E−07	0.018	1.18
	8	138,916,595	0.9	BLINK	5.67E−07	0.32	0.32
	11	49,075,590	0.7	BLINK	9.07E−07	0.076	0.62
		49,075,590	0.9	BLINK	2.51E−07	0.076	0.61
		278,897,832	0.9	BLINK	1.18E−06	0.005	2.77
		278,897,885	0.9	BLINK	1.18E−06	0.005	2.77
CTI	4	212,253,471	0.7	BLINK	3.97E−07	0.099	0.41
		212,253,471	0.9	BLINK	7.37E−08	0.099	0.41
	7	226,350,056	0.9	BLINK	1.89E−06	0.206	−0.3
		249,568,496	0.7	BLINK	5.64E−07	0.339	−0.26
		249,568,525	0.7	BLINK	4.08E−07	0.336	−0.26
		250,892,132	0.9	BLINK	9.71E−07	0.018	0.82
		250,896,197	0.9	BLINK	9.71E−07	0.018	0.82
	11	49,075,590	0.9	BLINK	6.14E−07	0.076	0.44

Surv, survival score; CTI, cold-tolerance index; Chr., chromosome; MAF, minor allele frequency.

Linkage disequilibrium (LD) analysis refined these associations to 10 distinct LD blocks. Among these, seven blocks were shared between Surv and the CTI, indicating a consistent set of genomic regions associated with cold tolerance across both traits. Three additional blocks were exclusive to the Surv on chromosomes 1, 8, and 11, suggesting that certain genetic factors specifically influence the discrete Surv rather than the continuous spectrum represented by the CTI (Table 3).

Table 3. Linkage disequilibrium blocks containing GWAS-detected significant SNP markers for Surv and CTI.

Trait	Chr.	Start (bp)	End (bp)	Size (kb)	Containing SNP
Surv	1	304,295,609	304,418,377	122.77	1_304295609
	8	138,819,376	138,916,595	97.22	8_138916595
	11	278,897,738	278,897,890	0.15	11_278897832, 11_278897885
Surv, CTI	4	212,214,305	212,253,515	39.21	4_212253471
	7	226,336,277	226,350,056	13.78	7_226350056
		249,499,169	249,568,499	69.33	7_249568496
		249,568,525	249,568,577	0.05	7_249568525
		250,881,412	250,896,195	14.78	7_250892132
		250,896,197	250,923,115	26.92	7_250896197
	11	48,943,763	49,075,590	131.83	11_49075590

Functional annotation of the genes within the 10 LD blocks identified above resulted in a list of 13 candidate genes, which we evaluated based on the effects of their different genotypes within the LD blocks on cold responses (Table 4). We focus here on three significant genes with notable allelic effects and known roles in stress acclimation. On chromosome 4, DEMF04G18820 was annotated as *tRNA METHYLTRANSFERASE 9* (*TRM9*), involved in methylation of the transfer RNA (tRNA) wobble base, allowing for the selective and efficient translation of specific stress-response proteins under stress [29, 30] (Fig. 4A). The variant SNP 4_212215160 in the 5' untranslated region (5' UTR) of *CaTRM9* showed a significant difference in the CTI between two homozygous genotypes among *C. annuum* accessions ($p = 0.015$), for which a statistically meaningful comparison was possible, confirming the influence of allelic variation on cold tolerance (Fig. 4B). On chromosome 7, DEMF07G34300 was annotated as $[Ca^{2+}]_{cyt}$-*ASSOCIATED PROTEIN KINASE 1* (*CAP1*), classified as a receptor-like kinase (RLK) gene; in Arabidopsis (*Arabidopsis thaliana*), CAP1 regulates Ca^{2+}-dependent cellular responses and the expression of its encoding gene is induced 3.2-fold upon cold stress [31] (Fig. 4C). The variant SNP 7_249568496 in the 3' UTR of *CaCAP1* also exhibited significant differences in the CTI between the two possible homozygous genotypes at this site ($p < 0.001$), supporting a role for this locus in stress response (Fig. 4D). On chromosome 11, DEMF11G09290 was annotated as *PROTEIN PHOSPHATASE*

2A-2 (*PP2A-2*), implicated in the modulation of stress signaling and maintenance of cell-membrane stability [32, 33] (Fig. 4E). The variant SNP 11_49075590 mapping to an intron of *CaPP2A-2* displayed significant differences in the CTI between the two possible homozygous genotypes (p = 0.046), supporting the contribution of this gene to cold tolerance (Fig. 4F). The genes exclusive to Surv were annotated as downstream effectors involved in cellular maintenance. On chromosome 1, DEMF01G43720 was annotated as *HEAT SHOCK PROTEIN 70-7* (*HSP70-7*) that functions in folding and degradation of damaged proteins [34]. On chromosome 8, DEMF08G09380 was annotated as *DIGALACTOSYLDIACYLGLYCEROL SYNTHASE 2* (*DGD2*), involved in maintaining the stability of the chloroplast membrane [35].

Table 4. Annotation of genes associated with Surv and CTI.

Trait	Locus identifier	Chr.	Start (bp)	End (bp)	Annotation
Surv	DEMF01G43720	1	304,404,633	304,406,225	Heat shock protein 70-7 (chloroplastic)
	DEMF08G09380	8	138,897,573	138,899,690	Digalactosyldiacylglycerol synthase 2 (chloroplastic)
	DEMF08G09410	8	138,909,783	138,912,343	NAD-dependent malic enzyme (mitochondrial)
	DEMF08G09430	8	138,821,086	138,826,489	Ribosomal protein L14
	DEMF08G09440	8	138,864,149	138,873,557	Short-chain dehydrogenase/reductase 2b
	DEMF08G09450	8	138,902,460	138,904,531	(+)-neomenthol dehydrogenase
	DEMF11G27250	11	278,897,655	278,900,028	ABC transporter G28
Surv, CTI	DEMF04G18820	4	212,212,492	212,216,012	tRNA methyltransferase 9
	DEMF07G34300	7	249,565,220	249,567,808	$[Ca^{2+}]$cyt-associated protein kinase 1 (At5g61350)
	DEMF07G34780	7	250,879,917	250,881,779	Cytochrome b561 (At4g18260)
	DEMF07G34800	7	250,892,189	250,895,346	Protein VACUOLESS 1
	DEMF07G34810	7	250,897,522	250,903,128	Protein VACUOLESS 1
	DEMF11G09290	11	49,074,974	49,081,203	Protein phosphatase 2A-2

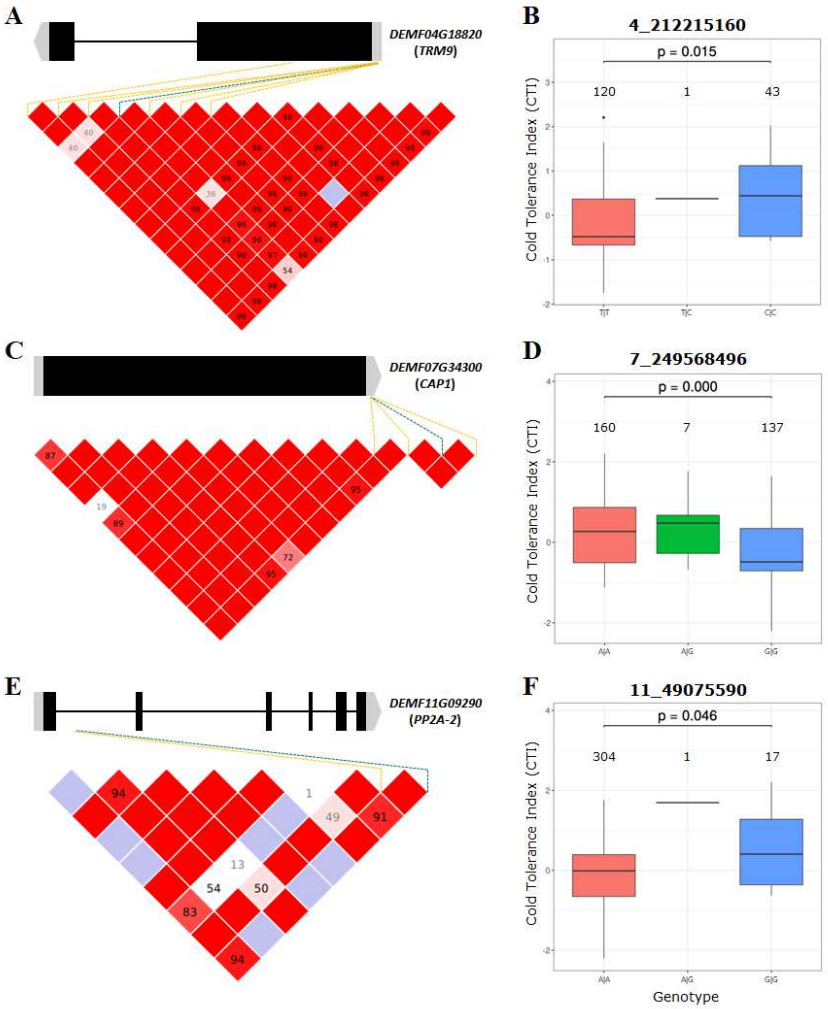

Figure 4. Fine mapping of candidate genes associated with cold tolerance and phenotypic differentiation by allele. (A, C, E) Diagrams showing the gene models (top) and linkage disequilibrium (LD) heatmaps (bottom) for candidate genes identified within significantly associated genomic regions. (A) DEMF04G18820 (*TRM9*) on chromosome 4. (C) DEMF07G34300 (*CAP1*) on chromosome 7. (E) DEMF11G09290 (*PP2A-2*) on chromosome 11. Black rectangles, exons; lines, introns. Yellow and green dashed lines connect the SNPs within LD blocks to their corresponding positions within the gene models; green dashed lines highlight the SNPs analyzed in panels B, D, and F. (B, D, F) Boxplots illustrating the effects of the different alleles at the significant SNPs on the CTI. In the boxplots, the center line indicates the median, the box limits represent the first and third quartiles, and the whiskers extend to 1.5 times the interquartile range. (B) *TRM9*-associated SNP 4_212215160. (D) *CAP1*-associated SNP 7_249568496. (F) *PP2A-2*-associated SNP 11_49075590. Statistical significance was determined using the Kruskal-Wallis test followed by Dunn's post-hoc test.

Optimization of genomic prediction accuracy using GWAS-informed markers

To evaluate the efficacy of a GWAS-informed marker selection strategy, we performed a ten-fold cross-validation on the core collection of 192 *C. annuum* accessions with up to 73,502 SNPs. We compared the predictive performance of 12 different GS models, encompassing penalized and mixed linear models, namely elastic net (EN), LASSO, ridge regression (RR), and RR-BLUP; the Bayesian linear models BayesB, BayesC, Bayesian LASSO (BLASSO), and Extended BLASSO (EBLASSO); and the non-linear models based on kernel-based approaches, such as reproducing kernel Hilbert space (RKHS) and support vector machine (SVM), and the two models random forest (RF) and XGBoost based on tree-based approaches (Table 5). The results consistently demonstrated that using subsets of top-ranked markers, which we selected based on their GWAS *p*-values, outperforms the use of the full marker set across all evaluated models (Fig. 5A). For both Surv and the CTI, prediction accuracies followed a bell-shaped trend: accuracy steadily rose as the number of top-ranked SNPs grew, peaking between 256 and 2,048 SNPs, before declining sharply as the inclusion of low-significance markers introduced environmental noise into the models.

323 **Table 5. Prediction accuracies of cross-validation for 12 genomic selection models with GWAS-based top-ranked marker sets.**

Trait	Num. of SNPs	Model											
		EN	LASSO	RR	RR-BLUP	BayesB	BayesC	BLASSO	EBLASSO	RKHS	RF	XGBoost	SVM
Surv	16	0.498	0.502	0.507	0.500	0.375	0.471	0.464	0.453	0.492	0.429	0.419	0.425
	32	0.503	0.505	0.514	0.519	0.378	0.495	0.481	0.480	0.516	0.451	0.466	0.449
	64	0.560	0.553	0.583	0.587	0.445	0.572	0.562	0.560	0.593	0.519	0.531	0.561
	128	0.564	0.558	0.586	0.578	0.435	0.540	0.548	0.543	0.581	0.482	0.499	0.554
	256	0.591	0.584	0.631	0.606	0.418	0.570	0.576	0.562	0.609	0.524	0.527	0.594
	512	0.680	0.672	0.698	0.671	0.471	0.631	0.644	0.635	0.670	0.561	0.566	0.637
	1,024	0.676	0.665	0.756	0.726	0.596	0.696	0.714	0.683	0.716	0.552	0.539	0.661
	2,048	0.715	0.707	0.749	0.741	0.722	0.712	0.735	0.713	0.729	0.510	0.481	0.660
	4,096	0.442	0.423	0.749	0.730	0.654	0.700	0.723	0.687	0.716	0.438	0.402	0.639
	8,192	0.121	0.080	0.666	0.708	0.660	0.657	0.714	0.645	0.691	0.331	0.270	0.589
	16,384	−0.030	−0.030	0.438	0.600	0.537	0.536	0.617	0.473	0.581	0.228	0.186	0.460
	32,768	−0.068	−0.075	0.171	0.101	0.129	0.153	0.148	0.110	0.093	0.085	0.082	0.108
	73,502	−0.089	−0.086	0.171	0.101	0.076	0.137	0.145	0.039	0.090	0.087	0.078	0.027
CTI	16	0.454	0.458	0.463	0.454	0.232	0.374	0.359	0.375	0.453	0.426	0.386	0.413
	32	0.496	0.498	0.498	0.483	0.253	0.416	0.392	0.374	0.485	0.411	0.444	0.430
	64	0.570	0.570	0.583	0.578	0.240	0.520	0.523	0.512	0.575	0.481	0.499	0.520
	128	0.684	0.681	0.695	0.691	-0.037	0.647	0.628	0.628	0.689	0.576	0.479	0.527
	256	0.702	0.698	0.739	0.707	0.249	0.674	0.642	0.651	0.702	0.606	0.500	0.574
	512	0.699	0.687	0.752	0.712	0.229	0.666	0.612	0.613	0.707	0.573	0.552	0.642
	1,024	0.702	0.696	0.780	0.762	0.362	0.700	0.577	0.450	0.750	0.567	0.528	0.685
	2,048	0.736	0.731	0.776	0.749	0.434	0.697	0.573	0.475	0.736	0.520	0.463	0.686
	4,096	0.465	0.442	0.751	0.740	0.332	0.675	0.474	0.420	0.726	0.458	0.368	0.659

8,192	0.141	0.111	0.478	0.709	0.141	0.583	0.174	0.260	0.693	0.358	0.273	0.610
16,384	−0.001	−0.001	0.451	0.626	−0.135	−0.307	−0.079	−0.377	0.610	0.264	0.179	0.508
32,768	−0.036	−0.066	0.202	0.141	−0.043	−0.143	−0.162	−0.060	0.134	0.124	0.088	0.199
73,502	−0.047	−0.07	0.203	0.141	−0.019	−0.073	−0.167	−0.022	0.134	0.121	0.097	0.145

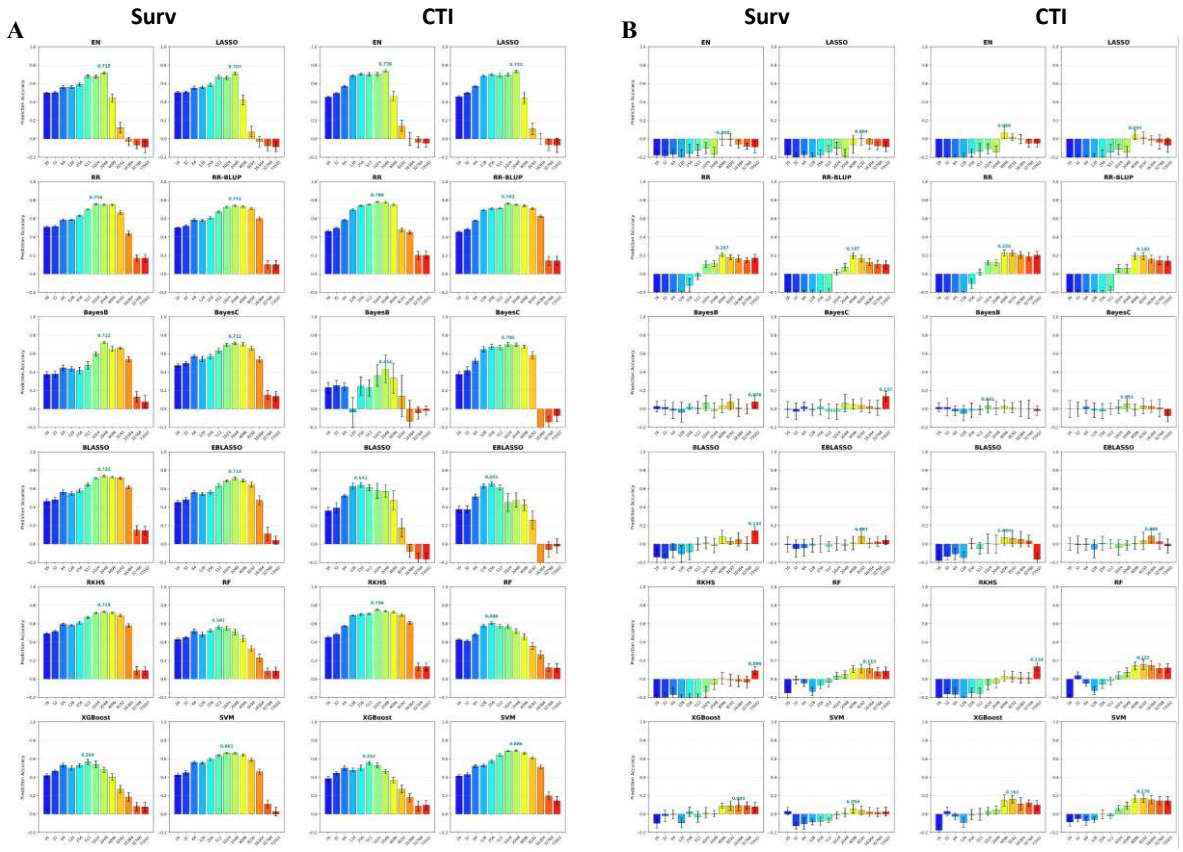

Figure 5. Cross-validation accuracies for 12 genomic selection models across various SNP panel sizes. (A) Results from cross-validation for Surv (first two columns) or CTI (third and fourth columns) when using GWAS-based top-ranked marker sets. (B) Results from cross-validation for Surv (first two columns) or CTI (third and fourth columns) when using randomly selected marker sets. The *x*-axes represent the number of markers used, from 16 to the full set of 73,502.

When using the full set of 73,502 SNPs, prediction accuracies were notably low, with the RR model yielding an average accuracy of only 0.203 for the CTI. By contrast, the GWAS-informed strategy achieved a significant enhancement in predictive power. For Surv, linear models reached high performance, with a peak accuracy of 0.756 for the RR model using the top 1,024 SNPs. The same RR model achieved its highest overall accuracy of 0.780 with the same subset of 1,024 SNPs for the CTI. Among non-linear models, kernel-based RKHS also demonstrated good performance, reaching a peak accuracy of 0.750 for the CTI, whereas tree-based models like RF and XGBoost showed comparatively

lower peak accuracies, ranging from 0.552 to 0.606.

A comparative analysis of model performance between the two traits revealed distinct preferences linked to trait architecture (Table 5). All penalized and mixed linear models reached higher peak accuracy when applied to the continuous CTI than to the discrete Surv (e.g., RR improved from 0.756 to 0.780, a statistically significant increase). This trend was mirrored in the non-linear models (RKHS, RF, and SVM). Conversely, Bayesian linear models consistently demonstrated superior performance for the discrete Surv trait but performed more poorly when applied to the CTI. Most notably, BayesB exhibited a sharp drop from 0.722 for Surv to 0.434 for the CTI. This divergence suggests that the sparse model assumptions of Bayesian methods, which emphasize a limited number of markers with relatively large effects while shrinking other marker effects to zero, may better align with the ordinal nature of the Surv scale, whereas linear and kernel-based models are better suited for capturing the polygenic architecture and continuous variation underlying the CTI.

To validate that the observed accuracy gains were due to the biological relevance of the GWAS-selected markers rather than a simple reduction in marker number, we compared the GWAS-informed sets against randomly selected marker subsets of identical sizes (Fig. 5B). For the CTI, while the RR model achieved a peak accuracy of 0.780 with 1,024 GWAS-ranked SNPs, the same model using 1,024 random SNPs yielded a low accuracy of only 0.122 (Table S2). Across all marker-set sizes and models, the random marker sets consistently produced accuracy values near zero or even negative, confirming that the GWAS-based prioritization successfully identified those genomic regions most critical for cold tolerance, effectively concentrating the genetic signal for high-precision genomic prediction.

Table S2. Prediction accuracies of cross-validation for 12 genomic selection models with randomly selected marker sets.

Trait	Num. of SNPs	Model											
		EN	LASSO	RR	RR-BLUP	BayesB	BayesC	BLASSO	EBLASSO	RKHS	RF	XGBoost	SVM
Surv	16	−0.178	−0.176	−0.234	−0.232	0.026	−0.007	−0.141	−0.008	−0.232	−0.151	−0.101	0.03
	32	−0.184	−0.202	−0.256	−0.223	0.017	−0.027	−0.156	−0.051	−0.208	−0.011	−0.019	−0.128
	64	−0.169	−0.171	−0.231	−0.228	−0.017	0.026	−0.070	−0.044	−0.169	−0.043	−0.005	−0.105
	128	−0.192	−0.207	−0.249	−0.232	−0.038	−0.009	−0.108	−0.015	−0.236	−0.135	−0.096	−0.085
	256	−0.158	−0.177	−0.123	−0.223	0.022	0.020	−0.085	−0.001	−0.234	−0.066	0.026	−0.078
	512	−0.122	−0.141	−0.025	−0.216	0.015	−0.028	−0.003	−0.026	−0.224	−0.037	−0.033	−0.068
	1,024	−0.101	−0.104	0.105	0.018	0.064	−0.031	0.010	0.002	−0.138	0.034	0.016	−0.012
	2,048	−0.165	−0.189	0.112	0.072	−0.019	0.063	−0.017	−0.015	−0.059	0.049	0.005	0.016
	4,096	−0.008	−0.064	0.207	0.197	0.033	0.050	0.078	0.011	0.007	0.108	0.089	0.059
	8,192	−0.009	0.004	0.18	0.166	0.072	0.041	0.029	0.083	−0.013	0.114	0.09	0.039
	16,384	−0.057	−0.057	0.166	0.127	0.009	0.026	0.048	0.012	−0.025	0.115	0.093	0.024
	32,768	−0.081	−0.078	0.149	0.106	−0.002	0.010	0.006	0.024	−0.032	0.082	0.092	0.018
	73,502	−0.089	−0.086	0.171	0.101	0.076	0.137	0.145	0.039	0.090	0.087	0.078	0.027
CTI	16	−0.223	−0.223	−0.225	−0.221	0.017	−0.002	−0.18	0.001	−0.25	−0.235	−0.177	−0.086
	32	−0.226	−0.221	−0.229	−0.221	0.019	−0.001	−0.134	−0.007	−0.161	0.037	0.03	−0.052
	64	−0.223	−0.224	−0.234	−0.221	−0.022	0.022	−0.108	−0.008	−0.166	−0.048	−0.025	−0.073
	128	−0.209	−0.21	−0.249	−0.221	−0.047	−0.014	−0.15	−0.055	−0.244	−0.127	−0.091	−0.06
	256	−0.153	−0.194	−0.107	−0.226	−0.017	−0.023	0.004	0.008	−0.15	−0.059	−0.007	−0.001
	512	−0.136	−0.143	0.017	−0.183	−0.012	0.002	−0.049	0.002	−0.155	−0.02	−0.002	−0.021
	1,024	−0.108	−0.116	0.122	0.061	0.031	0.021	0.005	−0.04	−0.061	0.039	0.03	0.062
	2,048	−0.142	−0.148	0.118	0.057	0.01	0.053	−0.004	−0.014	−0.043	0.073	0.046	0.09

4,096	0.068	0.044	0.225	0.193	0.028	−0.008	0.069	0.008	0.03	0.141	0.148	0.167
8,192	0.018	0.011	0.225	0.193	0.012	0.031	0.062	0.033	0.023	0.155	0.162	0.17
16,384	−0.003	−0.017	0.205	0.161	0	0.026	0.052	0.086	0.012	0.147	0.111	0.156
32,768	−0.048	−0.035	0.188	0.143	−0.003	0.008	0.032	0.025	0.011	0.118	0.12	0.145
73,502	−0.047	−0.07	0.203	0.141	−0.019	−0.073	−0.167	−0.022	0.134	0.121	0.097	0.145

Discussion

In this study, we implemented a GWAS-informed GS strategy at the yield-critical adult stage. By utilizing GWAS as a statistical filter to reduce noise and enhance accuracy, a method previously validated for capsaicinoid content [21] in pepper, we enhanced phenotypic resolution by using the continuous CTI and applied advanced genomic prediction to provide a robust framework for pepper breeding in an increasingly variable climate.

Phenotypic evaluation at the adult stage is limited by discrete visual scoring such as with Surv, which obscures the continuous spectrum of physiological stress responses. Our results confirmed this limitation, as Surv exhibited a stepped, non-normal distribution with most accessions clustered within a small range of values (Fig. 1B). To address this limitation, we developed the CTI by integrating Surv and the three agronomic traits PW, ST, and MFR (Fig. 1A). This integration converted discrete phenotypic data into a continuous, close to normal distribution, spreading clustered accessions across a wider phenotypic range (Fig. 1C). The CTI resolves the phenotypic compression inherent to the Surv scale, where accessions with the same Surv value differ in their recovery from cold stress and growth potential. While Surv reflects the extent of cold-induced damage, PW and ST quantify biomass accumulation and structural integrity [36, 37]; MFR is indicative of reproductive resilience at low temperatures [38]. These traits provide the physiological basis to disperse the accessions clustered by the Surv scale. The CTI therefore adds variance not captured during visual scoring for Surv, achieves the resolution necessary to dissect the polygenic architecture of cold tolerance, and satisfies the statistical assumptions required for genomic models.

Characterizing the genetic architecture of cold tolerance is a prerequisite for optimizing prediction models. SNP-based heritability analysis confirmed that Surv (h^2 = 0.53) and the CTI (h^2 = 0.55) are under genetic control. Despite environmental influence, significant SNPs associated with each trait were distributed across multiple chromosomes (Fig. 2). For Surv, chromosome 4 captured the highest proportion of genetic variance, followed by chromosome 3. For the CTI, chromosomes 3 and 4 showed substantial and nearly equal contributions to the total variance. Comparing this distribution with

GWAS results provides insight into the underlying genetic architecture (Table 2 and Figs. 2, 3). Chromosome 4 contributed to the genetic variance and showed significant associations for both traits. By contrast, chromosome 3 appeared to explain a substantial proportion of the variance but lacked significant SNPs by GWAS. While we detected significant genomic regions for Surv on chromosomes 1, 4, 6, 7, 8, and 11, signals for the CTI were concentrated on chromosomes 4, 7, and 11 (Table 2). Chromosome 3 showed high SNP-based heritability, but no significant GWAS signals. This suggests that many minor-effect variants on this chromosome contribute to cold tolerance, although the effect of each SNP is too small to be detected by GWAS. This observation indicates that the cold-tolerance trait is governed by the cumulative action of minor-effect genes that individually fall below the detection threshold but collectively drive phenotypic variation [5].

The high accuracies of Bayesian models for Surv (r = 0.712 to 0.735) may present a statistical paradox (Table 5 and Fig. 5A). Bayesian models are designed to capture a limited number of major-effect loci, assuming a sparse genetic architecture. Such models tend to underperform for polygenic traits governed by numerous minor-effect genes. The high accuracies for Surv may stem from the discrete nature of the visual scoring system, which forces the continuous spectrum of cold tolerance into an ordinal scale (Fig. 1). This discretization compresses polygenic variation into a simplified structure that aligns with the sparsity assumptions of Bayesian models. This interpretation is supported by the CTI results. When the trait was represented on a continuous scale, the performance of Bayesian models dropped to r = 0.434 to 0.700 (Table 5 and Fig. 5A). This decline in performance aligns with the theoretical expectation that Bayesian models fail to capture the minor effects characteristic of polygenic traits. The reduction in accuracy suggests that the previous high performance for Surv was a result of the discrete data structure rather than a reflection of the genetic architecture of the cold-tolerance trait.

Functional annotation of the candidate genes identified in this study aligned with abiotic stress–response mechanisms reported in previous studies (Table 4). The identification of a significant SNP associated with *TRM9* highlights the role of translational control in stress acclimation. TRM9 facilitates

stress acclimation by catalyzing the formation of 5-methoxycarbonylmethyluridine (mcm5U) at the tRNA wobble position (U34), thereby enabling the selective translation of stress-response mRNAs which, in contrast to regular transcripts, are uniquely enriched with AGA codons to prioritize their synthesis under environmental pressure [29, 30]. The genotype at the 5' UTR variant SNP 4_212215160 was associated with phenotypic differences in the CTI ($p = 0.015$), suggesting that polymorphisms in this regulatory region fine-tune translation efficiency or transcript stability (Fig. 4A, B). CAP1 is a calcium-dependent signaling component. In Arabidopsis, the expression of this RLK gene shows a 3.2-fold induction in response to cold stress [31]. RLKs perceive extracellular signals and initiate downstream phosphorylation cascades [39], which subsequently trigger or modulate intracellular cold-induced calcium signatures in pepper plants. The genotype at the 3' UTR variant SNP 7_249568496 was also associated with phenotypic differences in the CTI ($p < 0.001$; Fig. 4C, D). As 3' UTRs are involved in post-transcriptional regulation, this allelic variation suggests modulation of mRNA stability or localization. PP2A-2 is also associated with stress responses. In wheat, PP2As enhance cell-membrane stability under cold stress [32], while in rice, they are involved in signal modulation derived from high salinity and the combined stresses of drought and heat [33]. The genotype at the intronic variant SNP 11_49075590 displayed phenotypic differences in the CTI ($p = 0.046$), suggesting that non-coding regulatory mechanisms optimize phosphatase activity for cellular homeostasis (Fig. 4E, F). Genes exclusively associated with Surv were enriched in downstream effector functions in cellular maintenance, such as protein quality control via *HSP70-7* [34] and stabilization of the chloroplast membrane via *DGD2* [35]. This functional distinction indicates that the CTI captures regulatory components of the cold-stress response, distinct from the downstream effectors identified by Surv alone.

We detected the significant SNPs identifying these candidate genes with the BLINK model, while in other models the same SNPs failed to surpass the significance threshold (Table 2 and Fig. S2). We attribute this discrepancy to the BLINK algorithm, which employs iterative marker selection based on LD to eliminate redundant markers [40]. Unlike standard MLM or CMLM, which may suffer from reduced statistical power due to over-penalization for population structure, BLINK concentrates the

statistical signal onto a single representative marker within an LD block. This concentration drove the *p*-values of the representative SNPs below the Bonferroni threshold in the GWAS. This finding suggests that for polygenic traits with dispersed signals, multi-locus models capable of resolving LD structures may offer greater potential for effective locus detection.

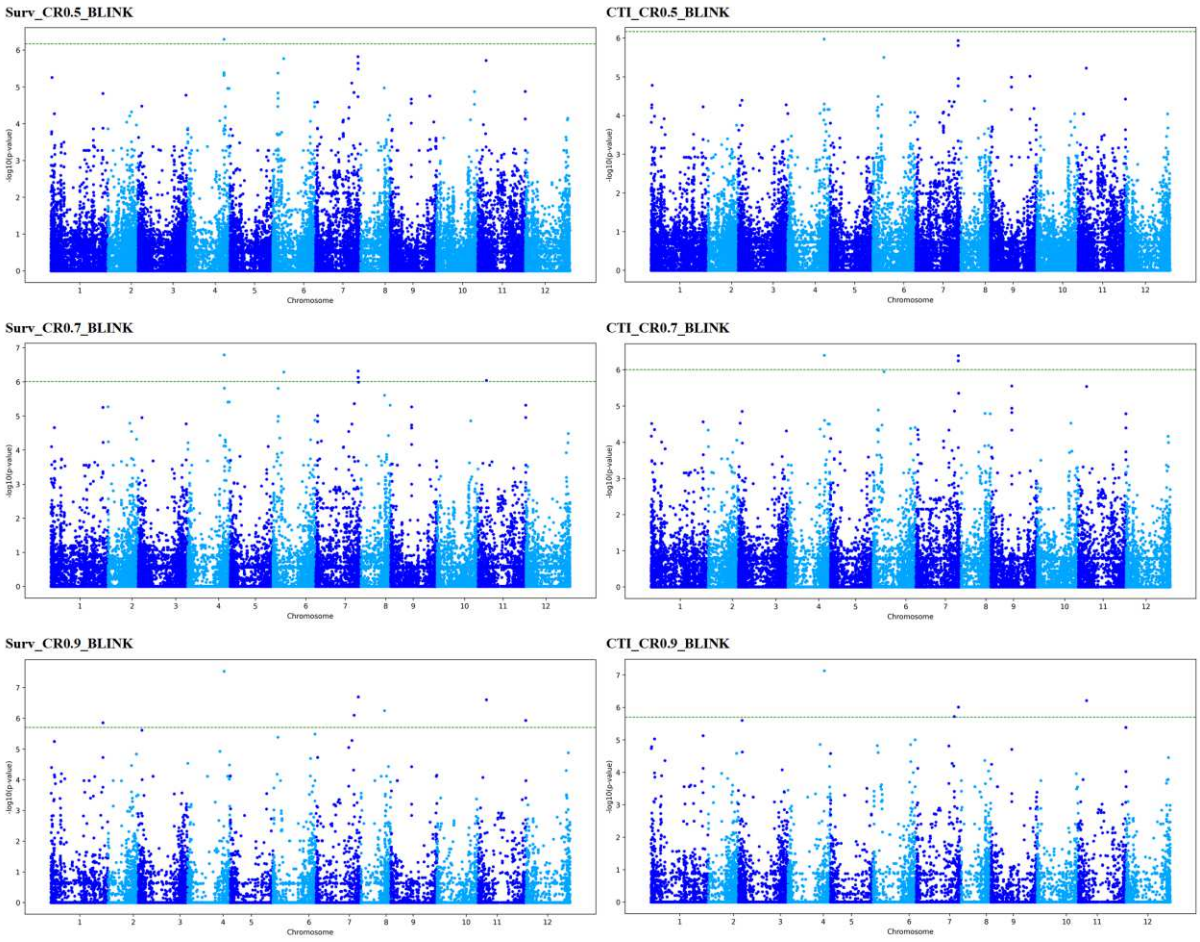

Fig. S2. Manhattan plots showing GWAS results for all traits and BLINK model. Manhattan plots display GWAS results obtained under different marker filtering conditions by call rate, with only the BLINK model shown. The x-axis represents the 12 chromosomes of the pepper genome, and the y-axis represents $-\log_{10}(p\text{-value})$ for SNP-trait associations. The horizontal dashed green line indicates the Bonferroni-corrected significance threshold.

The polygenic architecture of cold tolerance provides a biological rationale for the results of GS. Cold tolerance appears to be governed by widespread small-effect loci across the genome. Additive linear models (EN, LASSO, RR, and RR-BLUP) utilizing information from all markers consistently outperformed models assuming sparse architecture (BayesB, BayesC, BLASSO, and EBLASSO; Table 5 and Fig. 5A). Our GWAS-informed GS strategy demonstrated efficacy across all models. Filtering markers based on GWAS significance prioritized trait-associated variants while reducing statistical noise. All 12 models showed improvements when using top-ranked subsets compared to the full marker set. The RR model accuracy rose from 0.203 with the full 73,502 marker set to 0.780 with the top 1,024 GWAS-ranked SNPs (Table 5 and Fig. 5A). A comparative analysis of model performance across the two traits highlights the consistency of the CTI in detecting genomic regions associated with cold tolerance (Fig. 5). While Bayesian models performed well for the discrete Surv due to the categorical features of the data, the continuous CTI produced accuracies that were comparable but more aligned with additive linear models. Specifically, the RR model maintained its predictive performance with an accuracy of 0.780 for CTI, whereas Bayesian models showed a decline on the continuous scale. This change in model behavior suggests that the CTI refined the polygenic continuity of the trait, enabling linear models to capture minor effects in a manner consistent with the underlying genetic architecture of cold tolerance.

Conclusions

Our study established a robust framework for improving cold tolerance in adult pepper plants by integrating a new phenotypic index with a GWAS-informed genomic selection strategy. The development of the CTI overcame many of the limitations stemming from discrete visual scoring, enabling the precise capture of polygenic variation and the identification of candidate genes such as *TRM9*, *CAP1*, and *PP2A-2*. Furthermore, by incorporating these GWAS-derived biological priors into prediction models, GWAS-informed marker set achieved a 3.8-fold higher prediction accuracy

compared to the full marker set, demonstrating that prioritizing biologically relevant markers is far more effective than increasing marker density alone. These findings underscore the potential of integrating physiological and genomic data to accelerate genetic gain for complex abiotic stress traits. Moving forward, future research should focus on validating these results across multi-environment trials to ensure phenotypic stability and functionally characterize the identified candidate genes. Ultimately, the high-value marker sets identified here provide a solid foundation for developing cost-effective, low-density genotyping platforms to facilitate the practical implementation of genomic selection in commercial pepper breeding programs.

Methods

Plant materials and experimental design

A core collection of 192 *Capsicum annuum* accessions was used as the training population in this study [18] (Table S1). These accessions were cultivated with three plants per accession at the breeding station of NH Nongwoo Seed Turkey (Antalya, Turkey) over two consecutive growing seasons (2023–2024 and 2024–2025). Seeds were sown in seedling tray at July, and seedlings were transplanted to soil in a plastic greenhouse where they were kept until the following May. To evaluate cold tolerance at the adult stage, plants were exposed to natural low-temperature stress during the winter period (December to February), with temperatures ranging between 10 °C and 15 °C.

Table S1. List of 192 *Capsicum annuum* accessions.

No.	IT number	Accession name
1	IT103410	Gyeonggiyangpyeong-1985-gochu103410
2	IT113635	Gwangyang-3 (landrace)
3	IT135659	Ulju (landrace)
4	IT135673	Jinju (landrace)
5	IT135674	Hwaseong (landrace)
6	IT158280	Szechwan 8
7	IT158281	Szechwan 9
8	IT158351	Kradee Kiew

9	IT158627	HDA273
10	IT158628	HDA295
11	IT158680	L.P.1
12	IT158708	C00559
13	IT158714	PANGALENGAN-1
14	IT158719	C00590
15	IT158720	C00595
16	IT163484	PI249635
17	IT163497	YU1014
18	IT163529	Joseon pepper
19	IT163532	Seodong
20	IT163536	YU1130
21	IT213877	UZB-GJG-1999-51
22	IT219844	PBC566 PeMV Tex#1 GH#21
23	IT221673	Super Hot
24	IT221677	Thai Hot
25	IT223677	KC 00003
26	IT223737	VP 55
27	IT236368	Hu-33 AVRCD94187
28	IT236399	99TW Yellow 3
29	IT236666	Ikdosan
30	IT240012	Chili bangi #3
31	IT241900	Mestniy
32	IT247232	P 088-39
33	IT248579	VP 50
34	IT248583	VP 66
35	IT250180	A9E0762
36	IT250212	PE02
37	K011815	24B-2-1-4-3-2-1
38	K011821	24B-2-1-4-3-4-3
39	IT32436	Jindo (Landrace)
40	IT32447	Namwon (Landrace)
41	IT158278	MC4
42	IT158662	K1
43	IT183332	Sinheung 83-2852
44	IT223684	KC 00016
45	IT223795	P 97633
46	IT229667	Chilly BSS 213 F2
47	IT230612	Chorbadjiyska
48	IT236359	Lac-301

49	IT236369	Sacheon 94187
50	IT236370	Szechwan 2
51	IT236398	PX23595(No.8)
52	IT236424	Jungangjongmyo-2000-5032
53	IT247235	P 088-44
54	IT224926	Unknown 43 (straight)
55	IT225019	Macskasarga
56	IT228634	9852-193 AVRDC 211
57	K153298	Riot
58	IT231159	9250
59	IT228983	Numex Mirasol
60	IT134926	Taltos
61	IT158710	C00562
62	IT32404	Hotopep(345)(42-1S)
63	IT158279	MC5
64	K003577	SVS Bangalor 1
65	IT218757	Jungangjongmyo-2000-6447
66	IT32406	Hungarian Wax(355)(44-1S)
67	IT32506	Wonsi -1984+pepper032506
68	IT136602	Yatsu Fusha
69	IT136633	CC582 / NPL-CGS-1986-22892
70	IT215521	Nawash shon
71	IT259112	CHN-WJG-1999-29
72	IT32385	Saegochu (99)(13-1S)
73	IT32387	Saegochu (125)(15-1X)
74	IT32439	Yangpyeong (Landrace)
75	IT32463	Sangju (Landrace)
76	IT158692	RED & GREEN P5-4
77	IT236730	GWANGJU
78	IT237567	Mori tf.
79	IT237617	A9E0122
80	IT261350	Da niujiao
81	IT261405	09G111
82	IT158610	Sweet Banana
83	IT236277	Banana Supreme Hybrid
84	IT236340	New Mexico
85	IT236341	H.Sweet Wax
86	IT247225	P 088-20
87	IT247226	P 088-21
88	IT260880	THA-Kasassert Univ.-15

89	IT158788	C00881
90	IT221686	Starburst_2
91	IT237634	A9E0211
92	IT240859	Riben chaotian shu
93	IT261138	Ostry mnogoletny
94	IT229107	CMV 980
95	IT231151	Ibam
96	IT221683	Twilight
97	IT231171	9516
98	IT134925	Budatetenyi
99	IT258933	Zolotoi yubileinii
100	IT260894	55-03-01
101	IT199414	Dar Tashkent
102	IT219863	CCA1007
103	IT225013	Swedish
104	IT229690	Malishok
105	IT235383	Zavolgski 55
106	IT235605	Moldova-118
107	IT236344	Agronomico NO.8
108	IT236689	WIR 1570
109	IT240154	Volzanin
110	IT242040	Alena
111	IT242202	Suigyoku 2
112	IT247159	A9E0129
113	IT250182	A9E0764
114	IT261417	AC 09-035
115	IT158393	VAR.KAPIA F1-96
116	IT223790	Corne ll14
117	IT229133	CMV 1078/1
118	IT229136	CMV 1105
119	IT247244	P 088-57
120	K100100	HRF
121	K136576	Avelar 2001
122	K158293	Spanish Piquillo
123	K136529	Chocolate Beauty
124	IT158411	C01291
125	IT225016	Tura
126	IT224995	Canada cheese
127	IT163489	PI257052
128	IT216415	Jingle Bells Hybrid

129	IT225015	Greygo
130	IT225028	Okrasne papricky
131	IT231180	Yellow mushroom (berry type)
132	IT231180	Yellow mushroom (pepper type)
133	IT240349	99-10-47
134	IT250191	AC 09-020
135	IT250226	UZB-HHS-2008-15
136	IT250232	Hai li 10
137	IT261314	Dari Tashkenta
138	IT261596	A9E0692
139	IT278485	A9E0109
140	IT224899	Yiselie hong fei luo qing jiao
141	IT236179	Uz000170
142	K136596	5502
143	IT236754	Edesalma
144	IT224978	TC06651
145	IT240016	Jalapeno Mucho Nacho
146	IT248587	VP 84
147	IT237079	Kapia
148	IT229152	Pen-60/2/19914/1
149	IT221686	Starburst_1
150	K136588	Filus blue
151	K136521	Hot Jalapeno
152	IT293808	CMV 988/1
153	K136563	Avella
154	IT158309	HDA210
155	CC3677	9177_1
156	IT201167	ACC175
157	IT231150	C04398-A
158	K136593	CV3
159	K253872	CV4
160	K253873	CV8
161	K253874	CV9
162	IT221672	Dempsey
163	IT158583	Perennial
164	IT250209	CM334
165	IT158331	ECW
166	K253875	Special
167	K253876	35009
168	K253877	DRB

169	IT246553	Yuwolcho
170	K253878	35001
171	K253879	Taean jaerae
172	K253880	RS203
173	IT270425	Long Sweet
174	K150013	YCM334
175	K253881	Bukang C line
176	K253883	VK-515R
177	IT238119	Takanotsume
178	K136548	Jupiter
179	K136578	Yolo Y
180	IT158317	PI 201234
181	K253885	MicroPep Red
182	K253886	MicroPep Yellow
183	K253887	Milyang C (Milyang K)
184	IT113726	Uiseong jaerae 1
185	IT136632	NPL-CGS-1986-22891
186	IT158314	HDA268 / 88401-0
187	IT158332	VC32
188	IT158460	C01582
189	IT213876	MK-4522
190	IT216362	UZB-UzRIPI-1998-42
191	IT218653	GHA-LDJ-1998-100
192	IT223623	RUS-KHC-1999-36

Phenotypic evaluation

Phenotypic data were collected for the primary cold-tolerance trait and 13 additional agronomic traits. Assessments were primarily conducted in February of each year when plants had reached the adult stage and had been fully exposed to cold stress, with the exception of cotyledon color (CC), which was recorded at the seedling stage.

The primary survival score (Surv) was assessed visually using a nine-point scoring scale to capture the physiological response to cold stress. On this scale, a score of 1 indicates complete death, while a score of 9 represents vigorous recovery with growth and fruit setting comparable to plants growing under normal conditions. Intermediate scores reflect varying degrees of damage: a score of 3

indicates severe damage characterized by growth cessation and abscission of flowers or fruits; a score of 5 denotes moderate damage with leaf chlorosis or wilting and some fruit abscission, with maintenance of normal vegetative growth; a score of 7 represents continuous growth with no abscission and only minor stress symptoms (Fig. S3). To provide a robust benchmark for evaluating these physiological responses, the commercial *C. annuum* cultivar 'Pusula' was used as a primary check cultivar across all replicates. 'Pusula' is not part of the core collection.

Agronomic traits were categorized into continuous and categorical variables. Continuous traits were plant height (PH), measured as the distance from the ground to the top of the plant; plant width (PW), defined as the widest diameter of the plant canopy; main stem length (MSL), measured from the ground to the first branching node; and stem thickness (ST), recorded as the diameter of the main stem using a caliper. Reproductive continuous traits consisted of the average length (FL) and width (FWd) of representative mature fruits, as well as fruit weight (FWg), which was calculated by dividing the total fruit yield by the number of harvested fruits. Additionally, the mature fruit ratio (MFR) was calculated as the ratio between the number of mature fruits and the total number of fruits set, serving as a specific indicator of reproductive resilience under cold stress.

Categorical traits were scored based on established visual morphological standards. Plant habit (PHa) was scored on a 4-point scale ranging from prostrate (1) to fasciculate (4). Stem pubescence (SP) was scored on a 4-point scale based on trichome density, and stem color (SC) was graded from green (1) to purple (4). Similarly, leaf color (LC) was assessed on a 5-point scale from pale green (1) to purple (5), and CC was scored as green (1), green with purple (2), or purple (3).

To obtain representative phenotypic values for genetic analysis, the raw data were processed according to trait categories. For categorical traits, including Surv and visual scores, the arithmetic mean across replicates and years was calculated. For continuous agronomic traits, their BLUPs were estimated using the R package 'lme4' [41], accounting for year and replication effects.

Phenotypic evaluation and CTI development

To overcome limitations from Surv and its discrete scale, a continuous cold-tolerance index (CTI) was developed, following an established framework for statistical indices when direct economic weights are not available [42]. Sub-traits (PW, ST, and MFR) were selected for inclusion only if they exhibited a significant Pearson's correlation coefficient in at least 40 out of 50 iterations during a five-fold cross-validation step, repeated ten times. In detail, the core collection was randomly divided into five subsets. Four of these subsets were used to calculate Pearson's correlation coefficients between Surv and each agronomic trait, and the statistical significance of the correlations was evaluated at $p < 0.05$. This five-fold cross-validation procedure was repeated 50 times using different random partitions. Traits that showed significant correlations with Surv in at least 80% (40 out of 50) of the iterations were considered to be robustly associated with Surv. The CTI was calculated by integrating Surv with a weighted sum of these sub-traits (S) through a standardized sum approach. The integration followed the formula:

$$CTI = \alpha \times Z_{\mathrm{Surv}} + (1 - \alpha) \times S$$

where Z_{Surv} is the standardized Surv score, and α is a balancing factor determined as $\alpha = 1 - |r_{(Surv,S)}|$. All traits were Z-score standardized to account for differences in measurement units and value ranges, ensuring that the CTI was not dominated by traits with larger numerical scales. The α value is designed to adjust the contribution of Surv and other supporting traits according to the degree of information overlap between Surv and the integrated supporting traits, avoiding redundancy while preserving complementary information. The weighted sum S was calculated as $S = \sum w_i z_i$, where w_i represents the Pearson's correlation coefficient between trait *i* and Surv, and z_i is the standardized value of trait *i*.

Genotyping and genome-wide association study (GWAS)

Genotypic data were obtained from a previously reported genotyping-by-sequencing (GBS) dataset [18]. High-quality SNPs were selected based on a minor allele frequency (MAF) > 0.05 and a call rate > 0.5, using PLINK 1.9 for quality control [43]. To assess the genetic reliability of the traits, SNP-based heritability (h^2) was estimated using GCTA [44] by applying the genome-based restricted maximum

likelihood (GREML) model to the genomic relationship matrix (GRM) made by '--make-grm' option. GWAS was conducted using the R package 'GAPIT 3' [45]. Seven models were tested: BLINK [40], FramCPU [46], MLMM [47], MLM [48], CMLM [49], GLM [45], and SUPER [50]. Among these, the BLINK model was primarily used for signal detection. The significance threshold was determined using a Bonferroni correction ($p < 0.05 / N$, with N being the number of SNPs used). LD blocks were defined using the confidence interval method [51].

To evaluate the phenotypic differentiation of cold tolerance among different genotypic groups, several statistical analyses were performed. A Kruskal-Wallis test was first conducted to detect overall differences among groups. Subsequently, Dunn's test was performed for pairwise comparisons using the R package 'FSA' [52]. For the functional characterization of identified candidate genes, their annotation was performed by using the sequences of their encoded proteins as queries against the Swiss-Prot database [53] using the BLASTP algorithm. Significant matches were filtered based on an E-value threshold of 1×10^{-5}, and biological functions were assigned based on the descriptions of the top-scoring hits with the highest percentage of identity.

Genomic selection (GS) and marker-set design

The prediction accuracy of genomic estimated breeding values (GEBVs) was evaluated using 12 GS models categorized into three groups. Parametric models were EN [54], LASSO [55], RR [56], and RR-BLUP [57]. Bayesian models were BayesB, BayesC [58], BLASSO [59], and EBLASSO [60]. Non-parametric models were RKHS [61], RF [62], XGBoost [63], and SVM [64]. These models were implemented using the R packages 'rrBLUP' [57], 'BGLR' [65], and 'VIGoR' [66] with default parameter settings. Performance was assessed using a 10-fold cross-validation scheme. In detail, the core collection was randomly divided into 10 subsets (folds). In each time, 9 folds were used to train the models, and the phenotype of the remaining fold was predicted using the trained models. This was repeated 10 times by rotating the test fold so that each fold was predicted once, resulting in a complete set of predictions for the whole core collection. We repeated the entire 10-fold cross-validation process

10 times in different random partitions to evaluate the stability of the prediction accuracy. Prediction accuracy was calculated as the Pearson's correlation coefficient between GEBVs and observed phenotypes. To evaluate the efficacy of GWAS-informed selection, marker sets of varying sizes (from 16 to full set of 73,502 SNPs) were composed using top-ranked SNPs based on GWAS *p*-values. These were compared against randomly selected marker sets of identical sizes from entire 73,502 markers to validate the biological signal captured by the GWAS-informed approach.

Data availability

The data supporting the results in this study are included in this article and its supplementary files, or are available from the corresponding author on reasonable request.

Abbreviations

- **BLASSO:** Bayesian LASSO
- **BLUP:** Best linear unbiased prediction
- **CAP1:** $[Ca^{2+}]_{cyt}$-associated protein kinase 1
- **CC:** Cotyledon color
- **CTI:** Cold-tolerance index
- **DGD2:** DIGALACTOSYLDIACYLGLYCEROL SYNTHASE 2
- **EBLASSO:** Extended Bayesian LASSO
- **EN:** Elastic net
- **FL:** Fruit length
- **FWd:** Fruit width
- **FWg:** Fruit weight
- **GEBV:** Genomic estimated breeding value
- **GREML:** Genome-based restricted maximum likelihood
- **GRM:** Genomic relationship matrix

- **GS:** Genomic selection
- **GWAS:** Genome-wide association study
- **HSP70-7:** HEAT SHOCK PROTEIN 70-7
- **LASSO:** Least absolute shrinkage and selection operator
- **LC:** Leaf color
- **LD:** Linkage disequilibrium
- **MAF:** Minor allele frequency
- **MAS:** Marker-assisted selection
- **MFR:** Mature fruit ratio
- **MSL:** Main stem length
- **PH:** Plant height
- **PHa:** Plant habit
- **PP2A-2:** PROTEIN PHOSPHATASE 2A-2
- **PW:** Plant width
- **QTL:** Quantitative trait locus
- **RF:** Random forest
- **RKHS:** Reproducing kernel Hilbert space
- **RLK:** RECEPTOR-LIKE KINASE
- **RR:** Ridge regression
- **SC:** Stem color
- **SP:** Stem pubescence
- **ST:** Stem thickness
- **Surv:** Survival score
- **SVM:** Support vector machine
- **TRM9:** tRNA METHYLTRANSFERASE 9

References

[1] IPCC. Food security. In: Climate Change and Land. 2022. p. 437–550.

[2] Hassan MA, Xiang C, Farooq M, Muhammad N, Yan Z, Hui X, et al. Cold stress in wheat: plant acclimation responses and management strategies. Front Plant Sci. 2021;12:676884.

[3] Yu M, Luobu Z, Zhuoga D, Wei X, Tang Y. Advances in plant response to low-temperature stress. Plant Growth Regul. 2025;105:167–185.

[4] Ou LJ, Wei G, Zhang ZQ, Dai XZ, Zou XX. Effects of low temperature and low irradiance on the physiological characteristics and related gene expression of different pepper species. Photosynthetica. 2015;53:85–94.

[5] Mitchell-Olds T, Schmitt J. Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. Nature. 2006;441:947–952.

[6] Ma Y, Dai X, Xu Y, Luo W, Zheng X, Zeng D, Pan Y, Lin X, Liu H, Zhang D, et al. COLD1 Confers Chilling Tolerance in Rice. Cell. 2015;160:1209–1221.

[7] Zhu Y, Zhu G, Xu R, Jiao Z, Yang J, Lin T, Wang Z, Huang S, Chong L, Zhu J. A natural promoter variation of SlBBX31 confers enhanced cold tolerance during tomato domestication. Plant Biotechnol J. 2023;21:1033–1043.

[8] Munns R, James RA, Xu B, Athman A, Conn SJ, Jordans C, Byrt CS, Hare RA, Tyerman SD, Tester M, et al. Wheat grain yield on saline soils is improved by an ancestral Na+ transporter gene. Nat Biotechnol. 2012;30:360–364.

[9] Wu J, Lawit SJ, Weers B, Sun J, Mongar N, van Hemert J, Melo R, Meng X, Rupe M, Clapp J, et al. Overexpression of zmm28 increases maize grain yield in the field. Proc Natl Acad Sci USA. 2019;116:23850–23858.

[10] Bernardo R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 2008;48:1649–1664.

[11] Ray D, Chatterjee N. Effect of non-normality and low count variants on cross-phenotype association tests in GWAS. Eur J Hum Genet. 2020;28:300–312.

[12] Yang M, Wen Y, Zheng J, Zhang J, Zhao T, Feng J. Improving power of genome-wide association studies via transforming ordinal phenotypes into continuous phenotypes. Front Plant Sci. 2023;14:1247181.

[13] Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–1829.

[14] Crossa J, Campos G de los, Pérez P, Gianola D, Burgueño J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010;186:713–724.

[15] Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, et al. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics. 2014;15:740.

[16] Beyene Y, Semagn K, Mugo S, Tarekegne A, Babu R, Meisel B, et al. Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. Crop Sci. 2015;55:154–163.

[17] Hong JP, Ro N, Lee HY, Kim GW, Kwon JK, Yamamoto E, Kang BC. Genomic Selection for Prediction of Fruit-Related Traits in Pepper (*Capsicum* spp.). Front Plant Sci. 2020;11:570871.

[18] Lee HY, Ro NY, Jeong HJ, Kwon JK, Jo J, Ha Y, et al. Genetic diversity and population structure analysis to construct a core collection from a large *Capsicum* germplasm. BMC Genet. 2016;17:142.

[19] Alemu A, Åstrand J, Montesinos-López OA, Isidro y Sánchez J, Fernández-Gónzalez J, Tadesse W, et al. Genomic selection in plant breeding: key factors shaping two decades of progress. Mol Plant. 2024;17:552–578.

[20] Pang Y, Wu Y, Liu C, Li W, St Amand P, Bernardo A, et al. High-resolution genome-wide association study and genomic prediction for disease resistance and cold tolerance in wheat. Theor Appl Genet. 2021;134:2857–2873.

[21] Kim GW, Hong JP, Lee HY, Kwon JK, Kim DA, Kang BC. Genomic selection with fixed-effect markers improves the prediction accuracy for capsaicinoid contents in *Capsicum annuum*.

691 Hortic Res. 2022;9:uhac204.
692 [22] Alemu A, Batista L, Singh PK, Ceplitis A, Chawade A. Haplotype-tagged SNPs improve
693 genomic prediction accuracy for Fusarium head blight resistance and yield-related traits in
694 wheat. Theor Appl Genet. 2023;136:92.
695 [23] Zeng Y, Xu X, Jiang J, Lin S, Fan Z, Meng Y, et al. Genome-wide association analysis and
696 genomic selection for leaf-related traits of maize. PLoS One. 2025;20:e0323140.
697 [24] Lopez-Cruz M, Olson E, Rovere G, Crossa J, Dreisigacker S, Mondal S, et al. Regularized
698 selection indices for breeding value prediction using hyper-spectral image data. Sci Rep.
699 2020;10:8195.
700 [25] Crossa J, Cerón-Rojas JJ, Montesinos-López A, Montesinos-López OA, Punzalan J, Famoso
701 A, et al. Evaluating the effectiveness of selection indices and their genomic prediction using
702 environmental and historical rice data. G3 (Bethesda). 2025;15:jkaf087.
703 [26] Shikha K, Madhumal Thayil V, Shahi JP, Zaidi PH, Seetharam K, Nair SK, et al. Genomic-
704 regions associated with cold stress tolerance in Asia-adapted tropical maize germplasm. Sci
705 Rep. 2023;13:6297.
706 [27] Shah LR, Ahmed N, Hussain K, Mansoor S, Khan T, Khan I, et al. Mapping phenotypic
707 performance and novel SNPs for cold tolerance in tomato (*Solanum lycopersicum*) genotypes
708 through GWAS and population genetics. BMC Genom Data. 2024;25:9.
709 [28] Xiao J, Wang D, Liang L, Xie M, Tang Y, Lai Y-S, et al. CaMYB80 enhances the cold tolerance
710 of pepper by directly targeting *CaPOA1*. Hortic Res. 2024;11:uhae219.
711 [29] Leihne V, Kirpekar F, Vågbø CB, van den Born E, Krokan HE, Grini PE, et al. Roles of Trm9-
712 and ALKBH8-like proteins in the formation of modified wobble uridines in *Arabidopsis* tRNA.
713 Nucleic Acids Res. 2011;39:7688–7701.
714 [30] Wang Y, Pang C, Li X, Hu Z, Lv Z, Zheng B, et al. Identification of tRNA nucleoside
715 modification genes critical for stress response and development in rice and *Arabidopsis*. BMC
716 Plant Biol. 2017;17:261.

[31] Chae L, Sudat S, Dudoit S, Zhu T, Luan S. Diverse transcriptional programs associated with environmental stress and hormones in the *Arabidopsis* receptor-like kinase gene family. Mol Plant. 2009;2:84–107.

[32] Xu C, Jing R, Mao X, Jia X, Chang X. A wheat (*Triticum aestivum*) protein phosphatase 2A catalytic subunit gene provides enhanced drought tolerance in tobacco. Ann Bot. 2007;99:439–450.

[33] Yu RMK, Wong MML, Jack RW, Kong RYC. Structure, evolution and expression of a second subfamily of protein phosphatase 2A catalytic subunit genes in the rice plant (*Oryza sativa* L.). Planta. 2005;222:757–768.

[34] Saibil H. Chaperone machines for protein folding, unfolding and disaggregation. Nat Rev Mol Cell Biol. 2013;14:630–642.

[35] Klaus D, Härtel H, Fitzpatrick LM, Froehlich JE, Hubert J, Benning C, et al. Digalactosyldiacylglycerol synthesis in chloroplasts of the *Arabidopsis dgd1* mutant. Plant Physiol. 2002;128:885–895.

[36] Genzel F, Dicke MD, Junker-Frohn LV, Neuwohner A, Thiele B, Putz A, et al. Impact of moderate cold and salt stress on the accumulation of antioxidant flavonoids in the leaves of two *Capsicum* cultivars. J Agric Food Chem. 2021;69:6431–6443.

[37] Xia K, Hu P, Wu G, He X, Li Y, Li Y, et al. Prohexadione calcium boosts pepper stem rigidity: coordinated regulation of hormonal crosstalk and lignin deposition drives xylem reinforcement. BMC Plant Biol. 2025;25:1225.

[38] Pressman E, Shaked R, Firon N. Exposing pepper plants to high day temperatures prevents the adverse low night temperature symptoms. Physiol Plant. 2006;126:618–626.

[39] Ye Y, Ding Y, Jiang Q, Wang F, Sun J, Zhu C. The role of receptor-like protein kinases (RLKs) in abiotic stress response in plants. Plant Cell Rep. 2017;36:235–242.

[40] Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions.

743 Gigascience. 2019;8:giy154.
744 [41] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J
745 Stat Softw. 2015;67:1–48.
746 [42] Pešek J, Baker RJ. Desired improvement in relation to selection indices. Can J Plant Sci.
747 1969;49:803–804.
748 [43] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool
749 set for whole-genome association and population-based linkage analyses. Am J Hum Genet.
750 2007;81:559–575.
751 [44] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait
752 analysis. Am J Hum Genet. 2011;88:76–82.
753 [45] Wang J, Zhang Z. GAPIT version 3: boosting power and accuracy for genomic association and
754 prediction. Genom Proteom Bioinform. 2021;19:629–640.
755 [46] Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect
756 models for powerful and efficient genome-wide association studies. PLoS Genet.
757 2016;12:e1005767.
758 [47] Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus
759 mixed-model approach for genome-wide association studies in structured populations. Nat
760 Genet. 2012;44:825–830.
761 [48] Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-
762 model method for association mapping that accounts for multiple levels of relatedness. Nat
763 Genet. 2006;38:203–208.
764 [49] Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model
765 approach adapted for genome-wide association studies. Nat Genet. 2010;42:355–360.
766 [50] Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. A SUPER powerful method for genome wide
767 association study. PLoS One. 2014;9:e107684.
768 [51] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of

haplotype blocks in the human genome. Science. 2002;296:2225–2229.

[52] Ogle DH, Doll JC, Wheeler AP, Dinno A. FSA: simple fisheries stock assessment methods. R package. 2015.

[53] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. Nucleic Acids Res. 2025;53:D609–D617.

[54] Zou H, Hastie T. Addendum: regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67:768–768.

[55] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996;58:267–288.

[56] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55.

[57] Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4:250–255.

[58] Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinform. 2011;12:186.

[59] Park T, Casella G. The bayesian Lasso. J Am Stat Assoc. 2008;103:681–686.

[60] Mutshinda CM, Sillanpää MJ. Extended bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics. 2010;186:1067–1075.

[61] Gianola D, van Kaam JBCHM. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178:2289–2303.

[62] Breiman L. Random forests. Mach Learn. 2001;45:5–32.

[63] 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. p. 785–794.

[64] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14:199–222.

[65] Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198:483–495.

[66] Onogi A, Arakawa A. An R package VIGoR for joint estimation of multiple linear learners with variational bayesian inference. Bioinformatics. 2022;38:3306–3309.

Acknowledgements

Not applicable.

Funding

This work was supported by a grant from the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean government (MSIT) (No. RS-2023-00227464, Development of new varieties breeding technology with AI for strengthening food sovereignty) and by a grant from the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET) through the Agriculture and Food Convergence Technologies Program for Research Manpower Development, funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) (No. RS-2024-00398300).

Author information

Authors and affiliations

Department of Agriculture, Forestry and Bioresources, Research Institute of Agriculture and Life Sciences, Plant Genomics and Breeding Institute, College of Agriculture and Life Sciences Seoul National University, Seoul, 08826, South Korea

Kyeongseok Lee, Geon Woo Kim, Jin-Kyung Kwon, and Byoung-Cheorl Kang

Breeding Research Center, Nongwoo Seed Turkey, Muratpaşa, 07230, Turkey

Hee-Jin Jeong

Breeding Research Center, Nongwoo Bio, Yeoju, 12655, South Korea

Hyeon-Seok Jeong

Contributions

KL, GK, H-JJ, H-SJ, and BK conceived and designed the study. KL and H-JJ conducted the phenotypic evaluation, with H-JJ specifically managing the plant cultivation in Antalya, Turkey. KL and GK analyzed the phenotypic and genomic data. KL drafted the original manuscript, which was critically reviewed and revised by KL, GK, JK, and BK. BK supervised the project and provided overall guidance. All authors read and approved the final manuscript.

Corresponding author

Correspondence should go to Byoung-Cheorl Kang (bk54@snu.ac.kr).

Ethics declarations

Ethics approval and consent to participate

Not applicable. This study complies with all relevant local and national regulations.

Consent for publication

Not applicable.

Competing interests

Not applicable.