Research Square

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

Interpretable machine Learning Links Wheat 55K SNP Variation to Breeding-trait Prediction and Candidate Locus Prioritization

Haifang Sun
Hebei Academy of Agriculture and Forestry Sciences
Liang Hou
Hebei Academy of Agriculture and Forestry Sciences
Hao Qi
Hebei Academy of Agriculture and Forestry Sciences
Xiaorui Guo
Hebei Academy of Agriculture and Forestry Sciences
Jianan Min
Hebei Academy of Agriculture and Forestry Sciences
Shenglin Hou
Hebei Academy of Agriculture and Forestry Sciences
Ying Wang
Hebei Academy of Agriculture and Forestry Sciences
Xinshi Zhang
Hebei Academy of Agriculture and Forestry Sciences
Zhehao Tian
Hebei Academy of Agriculture and Forestry Sciences
Donghui Zhang
China Academy of Space Technology
Rui Wang
Inner Mongolia Academy of Science and Technology
Zhaoli An
Hebei Academy of Agriculture and Forestry Sciences
Xiaoliu Zheng
Hebei Academy of Agriculture and Forestry Sciences
Liangjie Lv
lvliangjie@haafs.org

Hebei Academy of Agriculture and Forestry Sciences

Article

Keywords: wheat, genomic prediction, SNP chip, machine learning, population structure, candidate SNP, spike number, plant height, yield

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10700600/v1

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

Additional Declarations: No competing interests reported.

Interpretable machine Learning Links Wheat 55K SNP Variation to Breeding-trait Prediction and Candidate Locus Prioritization

Haifang Sun[1], Liang Hou[1], Hao Qi[2], Xiaorui Guo[3], Jianan Min[1], Shenglin Hou[4], Ying Wang[1], Xinshi Zhang[1], Zhehao Tian[1], Donghui Zhang[5], Ruili Wang[6], Zhaoli An[1], Xiaoliu Zheng[1*], Liangjie Lv[3*]

[1]Institute of Agricultural Information and Economy, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, 050051, China;

[2]Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, 050051, China;

[3]Institute of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, 050035, China

[4]Institute of Agricultural Resources and Environment of Hebei Academy of Agriculture and Forestry Sciences, Shijiazhuang, 050051, China

[5]Institute of Remote Sensing Satellite, China Academy of Space Technology, Beijing, China

[6]Inner Mongolia Academy of Science and Technology, Hohhot, Inner Mongolia, China

*** Correspondence:**

Liangjie Lv lvliangjie@haafs.org

Xiaoliu Zheng zhengxiaoliu@haafs.org

Keywords: wheat; genomic prediction; SNP chip; machine learning; population structure; candidate SNP; spike number; plant height; yield

Abstract

Genomic prediction provides a route to accelerate wheat improvement by estimating breeding values directly from dense molecular markers, yet its practical value depends on both predictive reliability and biological interpretability. In many applied breeding populations, sample sizes are modest, phenotyping is limited to a small number of agronomically important traits, and marker effects are difficult to interpret beyond black-box prediction. Here, we developed an interpretable genomic prediction framework for three wheat breeding traits, plant height, plot yield and spike number per mu, using 55K SNP array data and matched field phenotypes. The analysis integrated SNP quality control, population structure assessment, repeated cross-validation, permutation testing, model comparison and candidate SNP prioritization within a unified workflow. After quality control, 2,310 SNPs from 193 genotyped wheat materials were retained, of which 180 materials had matched phenotypic records. Population structure analysis revealed substantial genomic stratification, with PC1 explaining 28.20% of SNP variation and showing a strong association with spike number per mu. Across 20 repeats of five-fold cross-validation, the best-performing models were trait-dependent. ExtraTrees achieved the highest accuracy for plant height and plot yield, with mean Pearson correlations of 0.3996 and 0.4103, respectively, whereas an RKHS-RBF model achieved high predictive accuracy for spike number per mu, with a mean Pearson correlation of 0.8721. Permutation tests confirmed that prediction was significantly better than random expectation for all three traits. Candidate SNP prioritisation identified lead markers on chr2B for plant height and plot yield, and a strong chr4B signal for spike number per mu. The chr4B lead SNP at 32,250,770 bp was located within TraesCS4B02G044500 and lay in a

candidate interval that physically overlapped a reported Rht-B1-associated selective sweep region, suggesting a possible relationship with plant architecture, spike-related traits or yield components. These results demonstrate that interpretable machine-learning-based genomic prediction can extract trait-relevant signals from a moderate-sized wheat SNP-chip dataset. The framework provides candidate markers for subsequent KASP assay development, expanded population validation and functional investigation, while also defining the limitations of inference from a single-environment breeding dataset

1 Introduction

Wheat is one of the most important staple crops worldwide, and continued improvement of wheat productivity remains essential for food security under changing production conditions. Breeding programmed must improve complex agronomic traits such as plant height, grain yield and yield components while maintaining adaptation to local environments. These traits are usually controlled by many loci with small to moderate effects, shaped by population structure and further modified by environmental variation. As a result, selection based only on observed field performance can be slow, costly and sensitive to year-specific or location-specific conditions (International Wheat Genome Sequencing Consortium, 2018; Cheng et al., 2024; Jiao et al., 2025).

Dense molecular markers provide an opportunity to accelerate breeding by linking genomic variation to phenotypic performance. SNP chips have become widely used in wheat because they offer stable marker sets, relatively low genotyping cost and genome-wide coverage suitable for applied breeding populations. The availability of the wheat reference genome has further improved the ability to place SNP markers into a physical genomic context and to interpret candidate loci near annotated genes. In practice, however, SNP-chip datasets from breeding programmes often contain missing genotypes, uneven marker distribution and moderate sample sizes. These features make it necessary to use modelling strategies that are robust, reproducible and interpretable (Rasheed et al., 2017; Sun et al., 2020; Walkowiak et al., 2020).

Genomic prediction estimates phenotypic performance or breeding value from genome-wide marker profiles. Classical models such as ridge regression, genomic best linear unbiased prediction and reproducing kernel Hilbert space regression have been widely used for genomic selection. More recently, machine-learning models, including tree ensembles, gradient boosting and neural networks, have been applied to capture nonlinear marker effects and interactions. These methods may improve prediction for traits with complex genetic architectures, but their performance depends strongly on trait type, population structure, sample size and validation design. Therefore, model evaluation must be based on repeated and properly separated cross-validation rather than a single random split (Meuwissen et al., 2001; Heslot et al., 2012; Crossa et al., 2017; Crossa et al., 2025; Montesinos-Lopez et al., 2021).

Prediction accuracy alone is not sufficient for breeding application. Breeders also need to know whether a model captures reproducible genomic signals, whether prediction exceeds random expectation, and which genomic regions contribute most strongly to trait variation. Interpretable genomic prediction therefore requires a workflow that connects model performance with candidate marker discovery. Such a workflow should include SNP quality control, population structure analysis, fair model comparison, permutation testing and genomic annotation of candidate loci. This is especially important for moderate-sized populations, where apparent signals may arise from structure or sampling effects if they are not examined carefully (Price et al., 2006; Semagn et al., 2021; Lundberg and Lee, 2017).

In this study, we developed an interpretable genomic prediction framework for wheat breeding traits using 55K SNP array data and matched phenotypes. We focused on three traits that are directly relevant to breeding practice: plant height, plot yield and spike number per mu. The workflow first standardised SNP coding and marker filtering, then evaluated population structure, compared multiple genomic prediction models under 20 repeats of five-fold cross-validation, tested whether predictive accuracy exceeded random expectation, and prioritised candidate SNPs by combining association strength, marker stability and model-derived importance (Rasheed and Xia, 2019; Crossa et al., 2025).

2 Materials and Methods

2.1 Study design

The study was designed to evaluate whether wheat 55K SNP-chip data could support both genomic prediction and interpretable candidate marker discovery for agronomic traits in a breeding population. The complete analytical framework included six modules: data integration, SNP recoding and quality control, population structure analysis, multi-model genomic prediction, permutation testing and candidate SNP annotation. **Figure 1** summarizes the workflow and shows how raw genotype and phenotype data were converted into predictive models and biologically interpretable candidate regions (Crossa et al., 2017; Rasheed and Xia, 2019).

The central input consisted of genotype data from wheat SNP-chip assays and matched phenotypic records for plant height, plot yield and spike number per mu. The primary outputs were cross-validated prediction accuracy, empirical significance from permutation testing, prioritised candidate SNPs and genomic annotation of candidate intervals.

2.2 Plant materials and phenotypic traits

The genotype dataset contained 193 wheat materials. Phenotypic records were available for 180 materials after matching genotype and phenotype identifiers. Three agronomic traits were used in this study: plant height, plot yield and spike number per mu. Plant height was recorded in centimetres. Plot yield was recorded as kg $plot^{-1}$. Spike number per mu was recorded as 10^4 spikes mu^{-1}.

For each trait, the number of available samples, mean, standard deviation, minimum, maximum and coefficient of variation were calculated. These statistics were used to evaluate whether the population contained sufficient phenotypic variation for genomic prediction. Pairwise trait correlations were also calculated to determine whether the three traits represented overlapping or distinct breeding targets.

2.3 SNP genotype recoding and quality control

Raw SNP-array genotypes were converted into numerical marker codes according to the reference and alternative alleles. For sample i and marker j, the recoded genotype was defined as:

$$x_{ij} = \begin{cases} 1, & \text{if the genotype was homozygous for the reference allele,} \\ -1, & \text{if the genotype was homozygous for the alternative allele,} \\ 0, & \text{if the genotype was heterozygous,} \\ 2, & \text{if the genotype was missing or undetermined.} \end{cases} \quad (1)$$

When the alternative allele was uncertain, homozygous genotypes different from the reference allele were treated as alternative homozygotes. Genotypes containing unexpected allele combinations were coded as heterozygous. This coding strategy followed the agreed preprocessing rule for converting

SNP-chip genotypes into the numerical format required by downstream prediction models (Wang et al., 2014; Rasheed et al., 2017; Sun et al., 2020).

Marker quality control was applied before all downstream analyses. For marker j, the missing rate was calculated as:

$$MR_j = \frac{n_{\text{missing},j}}{n} \tag{2}$$

where $n_{\text{missing},j}$ is the number of missing or undetermined calls and n is the number of genotyped materials. The minor allele frequency was calculated as:

$$MAF_j = min(p_j, 1 - p_j) \tag{3}$$

where p_j is the estimated allele frequency at marker j. SNPs were retained if they satisfied three criteria: missing rate ≤ 0.20, minor allele frequency ≥ 0.05 and valid genotype count ≥ 120. This filtering retained 2,310 SNPs from 5,589 available markers. Within each cross-validation split, missing marker values were imputed using training-set marker means to avoid information leakage into the held-out test set (Xiong et al., 2021; Mo et al., 2021; Fan et al., 2022).

2.4 Population structure analysis

Population structure was evaluated by principal component analysis based on the quality-controlled SNP matrix. Missing genotypes were handled during preprocessing before PCA. The first five principal components were retained to quantify major axes of genomic variation. The percentage of SNP variation explained by each principal component was calculated (Price et al., 2006; Semagn et al., 2021; Ali et al., 2022).

To evaluate whether population structure was related to agronomic traits, Pearson correlations were calculated between the leading principal components and each phenotype. This analysis was used to interpret model performance and candidate SNP discovery, especially for traits in which prediction accuracy could partly reflect structured genomic differences among materials.

2.5 Genomic prediction models

We compared models representing conventional genomic prediction, kernel learning, ensemble learning, gradient boosting and neural-network modelling. The baseline model predicted the training-set mean. Conventional and kernel models included ridge regression, genomic best linear unbiased prediction and a radial basis function reproducing kernel Hilbert space model. The radial basis function kernel was defined as: (Meuwissen et al., 2001; Heslot et al., 2012; Crossa et al., 2017)

$$K(x_i, x_{i'}) = \exp(-\gamma \|x_i - x_{i'}\|^2) \tag{4}$$

Where x_i and $x_{i'}$ are marker vectors for two materials and γ controls the kernel width.

Tree-based and boosting models included ExtraTrees, XGBoost, LightGBM and CatBoost. These models were included because they can capture nonlinear marker effects and higher-order interactions without requiring explicit interaction terms. A multilayer perceptron and a one-dimensional SNP

convolutional neural network were also evaluated to test whether neural-network models improved prediction in this moderate-sized dataset. A mean baseline model was included as a reference for non-informative prediction (Geurts et al., 2006; Chen and Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018; Montesinos-Lopez et al., 2021).

2.6 Cross-validation and model evaluation

Prediction performance was evaluated using 20 repeats of five-fold cross-validation, producing 100 train-test splits for each trait. In each split, marker selection, missing-value imputation, marker scaling and model fitting were performed only within the training set. The trained model was then applied to the held-out test set (Crossa et al., 2017; Vitale et al., 2025).

Model performance was evaluated using Pearson correlation, Spearman correlation, root mean squared error, mean absolute error, coefficient of determination and top 10% hit rate. Pearson correlation was calculated as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})\left(\hat{y}_i - \bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}} \quad (5)$$

where y_i is the observed phenotype and $\hat{y}_i$ is the predicted phenotype.

The root mean squared error was calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

The mean absolute error was calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

The coefficient of determination was calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

The top 10% hit rate was defined as:

$$H_{10} = \frac{\left|T_{\text{obs}} \cap T_{\text{pred}}\right|}{|T_{\text{obs}}|} \quad (9)$$

where T_{obs} is the set of top-ranked materials according to observed phenotypes and T_{pred} is the corresponding set according to predicted values. Lower values were treated as favourable for plant height, whereas higher values were treated as favourable for plot yield and spike number.

2.7 Permutation test

Phenotype-label permutation tests were used to determine whether the selected model for each trait exceeded random expectation. Phenotype values were randomly permuted 200 times while the genotype matrix remained unchanged. For each permutation, the same five-fold cross-validation procedure was applied and the mean Pearson correlation was recorded. The empirical P value was calculated as:

$$P_{\text{emp}} = \frac{1 + \sum_{b=1}^{B} I\left(r_b \geq r_{\text{obs}}\right)}{1 + B} \tag{10}$$

where r_{obs} is the observed prediction accuracy, r_b is the prediction accuracy from permutation b, B is the number of permutations, and $I(\cdot)$ is the indicator function.

2.8 Candidate SNP prioritization

Candidate SNPs were prioritised by integrating marker-trait association strength, false discovery rate support, fold-wise marker selection stability and model-based feature importance. For marker j, four component scores were calculated and min-max normalised within each trait:

$$Z_{cj} = \frac{C_{cj} - \min(C_c)}{\max(C_c) - \min(C_c)} \tag{11}$$

where C_{cj} is the raw value of component c for marker j. The stable marker score was then calculated as:

$$SMS_j = 0.35A_j + 0.20F_j + 0.25S_j + 0.20I_j \tag{12}$$

where A_j is the normalised association score based on $|r|$, F_j is the normalised false-discovery support based on $-\log_{10}(q)$, S_j is the normalised selection stability score, and I_j is the normalised feature-importance score. The importance component integrated ElasticNet coefficient magnitude, ExtraTrees importance, XGBoost importance, LightGBM importance, CatBoost importance and LightGBM mean absolute SHAP value. Higher stable marker scores indicated stronger integrated evidence (Benjamini and Hochberg, 1995; Lundberg and Lee, 2017).

Lead SNPs were identified for each trait based on the stable marker score and supporting statistical evidence. Candidate intervals were defined around clusters of high-priority SNPs on the same chromosome.

2.9 Genomic annotation and literature comparison

Candidate SNPs were mapped to their physical positions on the wheat reference genome. Nearby or overlapping genes were identified for lead SNPs and candidate intervals. For each trait, lead SNPs, candidate intervals and adjacent annotated genes were summarized. Candidate intervals were then compared with reported wheat genomic regions associated with plant height, architecture, yield components or domestication and breeding selection (International Wheat Genome Sequencing Consortium, 2018; Walkowiak et al., 2020; Jiao et al., 2025).

Particular attention was given to the chr4B candidate interval for spike number per mu because its lead SNP was located within an annotated gene and the interval physically overlapped a reported Rht-B1-associated selective sweep region. This overlap was interpreted cautiously as positional support for a

potentially relevant genomic region, not as proof that the lead SNP is Rht-B1 or that it has a confirmed functional effect (Semagn et al., 2021; Zhou et al., 2023).

2.10 Software and reproducibility

All analyses were implemented in Python. Data processing, model training, cross-validation, permutation testing, candidate SNP prioritisation and figure generation were performed using scripted workflows. The same quality-controlled SNP matrix and matched phenotype table were used consistently across all analyses to ensure reproducibility (Pedregosa et al., 2011).

3 Results and Analysis

3.1 An interpretable genomic prediction workflow links data quality, prediction and candidate-locus interpretation

We first assembled an interpretable genomic prediction framework that connected SNP-chip preprocessing, model evaluation and candidate locus prioritisation into a single evidence chain (Fig. 1a). The framework started from 193 genotyped wheat materials and retained 180 genotype-phenotype matched materials after sample integration. SNP markers were recoded into numerical genotype states, filtered by quality-control criteria and then used for population structure analysis, repeated genomic prediction, permutation-based validation and genomic interpretation.

Fig. 1b summarises the dataset retained for downstream analysis, with exact sample and marker counts reported in **Table 1**. After quality control, 2,310 of 5,589 SNPs were retained, corresponding to 41.3% of the original marker set, and 180 genotyped materials were matched with phenotypic records. The three target traits differed in phenotypic dispersion, with spike number per mu showing the largest coefficient of variation (CV = 12.9%), followed by plot yield (CV = 6.7%) and plant height (CV = 5.2%).

The selected trait-specific models are summarised in **Figure 1**c and **Table 2**. ExtraTrees achieved the highest prediction accuracy for plant height and plot yield, with mean Pearson correlations of 0.400 and 0.410, respectively. RKHS-RBF achieved the highest accuracy for spike number per mu, with a mean Pearson correlation of 0.872. Permutation testing further showed that all three selected models performed above random phenotype-label assignments.

3.2 Data overview and SNP quality control

The raw SNP dataset contained 5,589 markers with variable missing rates. After applying marker quality control, 2,310 SNPs were retained for downstream analysis (**Figure 2**a). The retained marker set represented 41.3% of the original SNPs and removed markers with excessive missingness. The minor allele frequency distribution after filtering showed that retained SNPs had allele-frequency variation suitable for genomic prediction, with a median MAF of 0.18 and a minimum retained MAF of 0.05 (**Figure 2**b).

The integrated dataset comprised 193 genotyped wheat materials and 180 samples with matched genotype and phenotype records (**Table 1**). From 5,589 raw SNP markers, 2,310 were retained after quality control and 3,279 were removed. SNP-level missing-rate and minor-allele-frequency distributions are shown in **Figure 2**a,b, whereas **Table 1** gives the exact dataset composition used for downstream analyses.

The retained SNPs covered the wheat genome, although marker density differed among chromosomes (**Figure 2**c). This distribution provided a genome-wide marker basis for population structure analysis, genomic prediction and candidate SNP prioritisation.

The three target traits showed continuous phenotypic variation (**Figure 2**d). Plant height was recorded for 180 samples (72.65 ± 3.81 cm), plot yield for 177 samples (10.60 ± 0.71 kg plot^{-1}) and spike number per mu for 180 samples (46.50 ± 5.99 × 10^4 spikes mu^{-1}). Pairwise trait correlations were low to moderate, with Pearson correlations of 0.20 between plant height and plot yield, 0.30 between plant height and spike number, and 0.03 between plot yield and spike number (**Figure 2**e).

3.3 Population structure revealed a major genomic axis associated with spike number

Principal component analysis of the quality-controlled SNP matrix revealed clear population structure among the 180 wheat materials (**Figure 3**a). PC1 explained 28.20% of genome-wide SNP variation, and PC2 explained 9.78%. The first two principal components separated the materials into broad genomic gradients rather than a single homogeneous cluster. When spike number per mu was mapped onto the PCA space, materials with higher spike number tended to occur along the positive PC1 direction. The first five principal components together explained 55.3% of SNP variation (**Figure 3**b). PC1 contributed the largest fraction of variation, followed by PC2, PC3, PC4 and PC5, which explained 9.78%, 6.91%, 5.42% and 4.97%, respectively. This pattern showed that a strong leading axis of genomic differentiation was present in the population.

PC-trait correlations further showed that population structure was trait dependent (**Figure 3**c). PC1 was strongly correlated with spike number per mu ($r = 0.773$, $P = 5.81 \times 10^{-37}$). PC1 was also correlated with plant height, although the magnitude was smaller ($r = 0.267$, $P = 2.98 \times 10^{-4}$). The association between PC1 and plot yield was weaker and marginal ($r = -0.147$, $P = 0.051$). Scatter plots of PC1 against the three traits confirmed this contrast: spike number per mu showed a pronounced PC1-aligned gradient, plant height showed a weaker positive trend and plot yield showed only a shallow negative trend (**Figure 3**d). These results established population structure as an important genomic background for interpreting prediction accuracy and candidate SNP discovery, particularly for spike number per mu.

3.4 Genomic prediction performance was trait dependent across model classes

We evaluated nine genomic prediction models using 20 × 5 repeated cross-validation. Prediction accuracy differed substantially among traits and model classes (**Figure 4**a). Plant height and plot yield were predicted with moderate accuracy, whereas spike number per mu showed high prediction accuracy across multiple models. For plant height, ExtraTrees achieved the highest mean Pearson correlation ($r = 0.400 \pm 0.108$). For plot yield, ExtraTrees was also the best-performing model ($r = 0.410 \pm 0.128$). For spike number per mu, RKHS-RBF achieved the highest accuracy ($r = 0.872 \pm 0.038$), followed closely by Ridge, GBLUP, ExtraTrees and CatBoost. The trait-specific best-model summary in **Figure 4**b highlights this difference between moderate prediction for plant height and plot yield and high prediction for spike number per mu.

The selected models and their exact performance metrics are reported in **Table 2**. Model errors relative to the mean baseline showed the same trait-dependent pattern (**Figure 4**d). For plant height, the best RMSE reduction was 7.5% with ExtraTrees. For plot yield, the best RMSE reduction was 8.1%, also with ExtraTrees. For spike number per mu, RKHS-RBF reduced RMSE by 50.7% relative to the mean baseline.

3.5 Observed and predicted values confirmed stronger prediction for spike number

We next examined observed-versus-predicted relationships for the best-performing trait-specific models. This analysis assessed whether cross-validated models captured genotype-level phenotypic ranking rather than only producing favourable average metrics.

For plant height, ExtraTrees produced a sample-level observed-predicted correlation of r = 0.411, with a cross-validation mean correlation of r = 0.400 ± 0.108 and RMSE = 3.53 cm (**Figure 5**a). Predictions followed the observed trend but were compressed toward the population mean, indicating moderate predictive resolution. For plot yield, ExtraTrees produced a sample-level correlation of r = 0.428, with a cross-validation mean correlation of r = 0.410 ± 0.128 and RMSE = 0.65 kg plot^{-1} (**Figure 5**b). The observed-versus-predicted pattern again showed a positive but moderate relationship.

Spike number per mu showed a much stronger observed-predicted relationship (**Figure 5**c). The RKHS-RBF model achieved a sample-level correlation of r = 0.877, with a cross-validation mean correlation of r = 0.872 ± 0.038 and RMSE = 2.95 × 10^4 spikes mu^{-1}. The predicted values closely followed the observed phenotypic gradient, although some shrinkage remained at the extremes. The residual distributions in **Figure 5** further showed that prediction errors were centred near zero for all three traits, with the strongest agreement observed for spike number per mu.

3.6 Permutation tests confirmed non-random genomic prediction

Permutation testing was used to determine whether observed prediction accuracy could arise from random phenotype-label assignments. For each trait, phenotype labels were randomly permuted 200 times while the genotype matrix was held fixed. The selected model for each trait was then evaluated against this null distribution. For plant height, ExtraTrees achieved significant non-random prediction, with an empirical permutation P value of 0.00498 (**Figure 6**a and **Table 2**). For plot yield, ExtraTrees also exceeded the permutation null expectation, with an empirical P value of 0.00995 (**Figure 6**b and **Table 2**). For spike number per mu, RKHS-RBF showed the strongest separation from the permutation null distribution, with an empirical P value of 0.00498 (**Figure 6**c and **Table 2**).

The distance between observed and permuted prediction accuracy further confirmed this pattern (**Figure 6**d). The observed accuracies were separated from their permutation null distributions by z = 3.35 for plant height, z = 3.40 for plot yield and z = 7.81 for spike number per mu. These results showed that the genomic prediction signals were not attributable to random phenotype-marker alignment.

3.7 Integrated marker evidence identified trait-specific candidate SNPs

Candidate SNP discovery combined single-marker association evidence, false-discovery correction, repeated selection stability and model-derived importance into a stable marker score. Genome-wide stability-selection profiles revealed trait-specific candidate regions across the wheat genome (**Figure 7**a). The leading plant-height signal was located on chr2B near 7.64 Mb. The leading plot-yield signal was also on chr2B near 18.94 Mb. The strongest spike-number signal was located on chr4B near 32.25 Mb.

The top-ranked SNPs showed distinct stability patterns for each trait (**Figure 7**b and **Table 3**). For plant height, the lead SNP was chr2B:7,644,927, with a stable marker score of 0.918, |r| = 0.342 and q = 2.05 × 10^{-3}. Several nearby chr2B markers also ranked highly, indicating a local candidate region around 7.64 Mb. For plot yield, the lead SNP was chr2B:18,937,487, with a stable marker score of 0.773, |r| = 0.309 and q = 3.07 × 10^{-3}. Additional high-ranking yield markers were observed on chr5D

and chr7B. For spike number per mu, the lead SNP was chr4B:32,250,770, with a stable marker score of 0.955, $|r| = 0.416$ and $q = 7.36 \times 10^{-6}$.

Candidate locus distributions showed that prioritised loci were not uniformly distributed across chromosomes (**Figure 7**c). Plant-height and plot-yield signals included chr2B lead regions, whereas spike-number signals included a high-priority chr4B interval. This trait-dependent distribution supported separate annotation of the lead SNPs rather than pooling all traits into a single candidate list.

3.8 Genomic annotation connected the chr4B spike-number signal to a candidate gene-level interval

We annotated candidate SNPs by physical position and compared priority intervals with literature-supported genomic regions. The most notable result was observed for spike number per mu on chromosome 4B. The candidate interval extended from chr4B:28,954,548 to chr4B:32,250,770, with the lead SNP located at chr4B:32,250,770. This interval physically overlapped a reported Rht-B1-related chromosome 4B selective sweep region extending from 30.3 to 34.2 Mb (**Figure 8**a). The shared overlap between the candidate interval and the literature region was 1.95 Mb.

Within the spike-number interval, multiple candidate SNPs and annotated genes were present (**Figure 8**b). The lead SNP had the highest stable marker score in this interval and was positioned at the distal end of the candidate region. Gene-track annotation showed that this SNP was located within an annotated gene-dense region. Gene-level mapping resolved chr4B:32,250,770 to TraesCS4B02G044500 (**Figure 8**c and **Table 3**). The lead SNP was located within the annotated gene coordinates and had a stable marker score of 0.955, a marker-trait correlation of 0.416 and a q value of 7.36×10^{-6}. The evidence summary in **Figure 8**d therefore connected the chr4B spike-number signal to a prioritised annotated interval for subsequent validation.

4 Discussion

4.1 Trait-dependent genomic predictability

This study developed an interpretable genomic prediction framework for wheat breeding traits by integrating SNP quality control, population structure analysis, repeated cross-validation, permutation testing and candidate SNP annotation. The framework produced significant predictive signals for plant height, plot yield and spike number per mu, but the strength of prediction differed markedly among traits. Plant height and plot yield showed moderate predictability, whereas spike number per mu showed high predictability across repeated validation. This conclusion was supported by the data-quality overview, model comparison, observed-versus-predicted analysis and permutation testing (Figs. 2, 4-6) (Crossa et al., 2017; Crossa et al., 2025).

The overall results show that SNP-chip data can support genomic prediction in a moderate-sized wheat breeding population when model evaluation and interpretation are treated as connected analytical steps. Rather than relying on a single prediction metric, the framework combined 20×5 repeated cross-validation, held-out prediction visualisation, empirical permutation testing and candidate locus annotation. This structure reduced the risk of over-interpreting model accuracy and provided a clearer route from prediction performance to biologically interpretable genomic regions.

4.2 Spike number per mu showed the strongest genomic signal

The high predictability of spike number per mu was supported by several convergent observations. First, spike number showed the largest phenotypic dispersion among the three traits, with $\mathrm{CV} = 12.9\%$, providing a wider phenotypic range for model learning (**Figure 2**d). Second, PC1 captured a major axis of genomic differentiation and was strongly correlated with spike number per mu ($r = 0.773,\ P = 5.81 \times 10^{-37}$), indicating that population structure contributed to the genomic signal for this trait (**Figure 3**). Third, RKHS-RBF consistently achieved the highest prediction accuracy for spike number per mu under 20×5 repeated cross-validation ($\bar{r} = 0.872 \pm 0.038$), and the observed-versus-predicted relationship showed close agreement between measured and predicted values (Figs. 4 and 5). Finally, permutation testing placed the observed accuracy far outside the random-label null distribution $\left(z = 7.81, P_{\mathrm{emp}} = 0.00498\right)$, confirming that the predictive signal was not attributable to random phenotype-marker alignment (**Figure 6**).

This result should be interpreted with both confidence and caution. The confidence comes from the consistency of evidence across phenotype distribution, population structure, cross-validation and permutation testing. The caution comes from the strong relationship between PC1 and spike number per mu. A high prediction accuracy for this trait may therefore reflect both marker-level trait information and structured genomic differentiation among materials. This does not invalidate the prediction result, but it means that independent population validation is required before the model is used as a general breeding predictor (Price et al., 2006; Semagn et al., 2021; Ali et al., 2022).

4.3 Model performance differed among traits

The different best-performing models across traits indicate that the genomic architecture and data structure were trait dependent. ExtraTrees performed best for plant height and plot yield, whereas RKHS-RBF performed best for spike number per mu. Tree-based ensemble models can capture nonlinear marker effects and local interactions while remaining relatively robust in moderate-sized datasets. This property may explain their advantage for plant height and plot yield, both of which showed moderate prediction accuracy and weaker PC1-aligned structure (Azodi et al., 2019; Crossa et al., 2025).

In contrast, the strong performance of RKHS-RBF for spike number per mu indicates that genome-wide similarity was highly informative for this trait. Kernel methods are well suited to capturing broad genomic relatedness, which is consistent with the strong PC1-associated phenotypic gradient observed for spike number per mu. The weaker performance of the SNP-1D CNN, especially for plot yield, is also informative. It suggests that the current sample size was insufficient for deep neural-network models to outperform simpler and more regularised approaches. For datasets of this scale, robust ensemble and kernel models may therefore be more appropriate than high-parameter deep-learning architectures (Montesinos-Lopez et al., 2021; Montesinos-Lopez et al., 2023; Togninalli et al., 2023; Montesinos-Lopez et al., 2024a; Kaushal et al., 2024).

4.4 Candidate SNPs linked prediction signals to genomic regions

The candidate SNP analysis provided a bridge between prediction accuracy and biological interpretation. Integrated marker evidence prioritised chr2B:7,644,927 for plant height, chr2B:18,937,487 for plot yield and chr4B:32,250,770 for spike number per mu (**Figure 7**). These candidate SNPs were supported by association strength, false-discovery evidence, selection stability and model-derived importance. The plant-height and plot-yield signals on chr2B indicate that this chromosome contributed to multiple agronomic traits in the present population (Rasheed and Xia, 2019; Fan et al., 2022).

Yield-associated signals were also observed on chr5D and chr7B, consistent with the distributed and quantitative nature of yield formation. This pattern agrees with the expectation that plot yield is controlled by multiple genomic regions and is sensitive to environmental and management effects. These loci should therefore be interpreted as candidate regions for marker development and validation, rather than deterministic predictors of breeding performance (Jackson et al., 2023; Vitale et al., 2025; Liu et al., 2025; Yannam et al., 2025).

4.5 The chr4B spike-number interval provides a priority target for validation

The chr4B spike-number interval provides the clearest example of how prediction and interpretation were connected in this study. The lead SNP, chr4B:32,250,770, had the highest stable marker score among the three lead markers ($\mathrm{SMS} = 0.955$), showed the strongest single-marker association ($|r| = 0.416,\ q = 7.36 \times 10^{-6}$), and was located within TraesCS4B02G044500. (Figs. 7 and 8) (Semagn et al., 2021; Zhou et al., 2023).

The broader candidate interval from 28.95 to 32.25 Mb physically overlapped a reported Rht-B1-related selective sweep region on chromosome 4B. This overlap provides positional support for the importance of this genomic region in plant architecture or yield-component variation. However, it does not demonstrate that the lead SNP is Rht-B1, nor does it establish a causal relationship between TraesCS4B02G044500 and spike number per mu. The result should be interpreted as locus-level evidence that prioritises the chr4B region for follow-up validation (Semagn et al., 2021; Xiong et al., 2021; Zhou et al., 2023).

4.6 Breeding implications

From a breeding perspective, the results support two complementary uses of SNP-chip data. The first is genomic prediction, where models such as ExtraTrees and RKHS-RBF can estimate trait performance from genome-wide marker profiles. The second is candidate marker prioritisation, where prediction-supported SNPs can be converted into breeder-friendly assays for independent validation (Rasheed and Xia, 2019; Crossa et al., 2025).

The candidate SNPs identified here, particularly those on chr2B and chr4B, may be useful for KASP marker development, marker-assisted preselection and expanded population testing. Their practical value will depend on whether the same marker-trait relationships are reproduced across additional germplasm, years and locations. The current results therefore provide a prioritised marker set for downstream breeding research, rather than a final marker panel for immediate deployment (Rasheed and Xia, 2019; Crossa et al., 2025).

4.7 Limitations and future directions

Several limitations should be considered when interpreting these findings. The matched population size was $n = 180$ for plant height and spike number per mu and $n = 177$ for plot yield. This scale is suitable for exploratory genomic prediction but remains limited for fine-mapping or deep-learning-based inference. The current analysis was also based on the available phenotypic dataset rather than a complete multi-year and multi-location trial series. Therefore, genotype-by-environment interaction could not be explicitly modelled (Jackson et al., 2023; Liu et al., 2025; Yannam et al., 2025; Kick et al., 2023; Jubair and Domaratzki, 2023).

In addition, the candidate genes were inferred from physical position and marker evidence, but functional validation was not performed. These limitations are especially important for plant height

and plot yield, which are more directly influenced by environmental and management conditions. Future work should extend the present framework to larger and more diverse wheat populations. Multi-environment phenotypes would allow the separation of stable genomic effects from environment-specific responses. Genotype-by-environment models could further improve prediction for traits such as yield. Finally, the lead SNPs identified in this study should be validated using independent populations, targeted genotyping and functional experiments (Jackson et al., 2023; Liu et al., 2025; Yannam et al., 2025).

5 Conclusion

This study established an interpretable genomic prediction framework for wheat plant height, plot yield and spike number per mu using 55K SNP-chip data. By combining SNP quality control, population structure analysis, 20×5 repeated cross-validation, permutation testing and candidate SNP annotation, the framework linked prediction performance with genomic interpretation. All three traits showed significant genomic predictability. ExtraTrees achieved the best performance for plant height and plot yield, whereas RKHS-RBF achieved the highest accuracy for spike number per mu. The strongest predictive signal was observed for spike number per mu, supported by its phenotypic variation, PC1-associated population structure, high cross-validated prediction accuracy and significant permutation-test result (Figs. 2-6).

Candidate SNP analysis identified trait-associated regions on chr2B, chr4B, chr5D and chr7B. The chr4B lead SNP for spike number per mu, chr4B:32,250,770, was located within TraesCS4B02G044500 and overlapped a reported Rht-B1-related chromosome 4B selective sweep interval at the locus level (Figs. 7 and 8). These results provide candidate markers and genomic regions for subsequent KASP assay development, multi-environment validation and functional investigation. The findings support the feasibility of using interpretable genomic prediction to extract breeding-relevant SNP signals from moderate-sized wheat SNP-chip datasets (Crossa et al., 2025; Gao et al., 2025). The conclusions remain bounded by the current population size, available phenotypic environments and lack of functional validation. Therefore, the candidate SNPs identified here should be regarded as prioritised targets for validation rather than confirmed causal loci or immediately deployable diagnostic markers.

Funding

This research was funded by Science and Technology Innovation Special Project of Hebei Academy of Agriculture and Forestry Sciences (2026KJCXZX-NXS-4), Hebei Academy of Agriculture and Forestry Sciences Basic Scientific Research Business Fund Youth Project (2025090103), Central funds project for guiding local technological development (254Z6303G), The earmarked fund for Hebei Agriculture Research System (HBCT2023010301), Funding for Talent Cultivation under the "333 Talent Project" (C2024082), and Natural Science Foundation of Inner Mongolia Autonomous Region Project (2024ZD18).

Institutional Review Board Statement

Not applicable for studies not involving humans or animals.

Informed Consent Statement

Not applicable for studies not involving humans.

Data Availability Statement

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Author contributions

H. S.: Conceptualization, Methodology, Writing – original draft, Visualization, Formal Analysis. L. H.: Investigation, Software, Data curation, Validation, Writing – review & editing. H. Q.: Investigation, Data curation, Formal Analysis, Writing – review & editing. X. G.: Software, Validation, Visualization, Writing – review & editing. J. M.: Investigation, Resources, Writing – review & editing. S. H.: Investigation, Data curation, Writing – review & editing. Y. W.: Investigation, Resources, Writing – review & editing. X. Z.: Formal Analysis, Validation, Writing – review & editing. Z. T.: Software, Validation, Writing – review & editing. D. Z.: Software, Data curation, Writing – review & editing, Validation, Visualization, Formal Analysis. R. W.: Validation, Writing – review & editing, Investigation, Resources, Software, Project administration, Supervision, Funding acquisition. Z. A.: Writing – review & editing, Validation. X. Z.: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – review & editing (Xiaoliu Zheng). L. Lv.: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Conflicts of Interest

The authors declare no conflict of interest.

References

Ali, M., Danting, S., Wang, J., Sadiq, H., Rasheed, A., He, Z., and Li, H. (2022). "Genetic diversity and selection signatures in synthetic-derived wheats and modern spring wheat." Frontiers in Plant Science 13, 877496. doi: 10.3389/fpls.2022.877496

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). "Benchmarking parametric and machine learning models for genomic prediction of complex traits." G3: Genes, Genomes, Genetics 9, 3691-3702. doi: 10.1534/g3.119.400498

Benjamini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." Journal of the Royal Statistical Society: Series B 57, 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi: 10.1145/2939672.2939785

Cheng, S., Feng, C., Wingen, L. U., Cheng, H., Riche, A. B., Jiang, M., Leverington-Waite, M., Huang, Z., Collier, S., Orford, S., et al. (2024). "Harnessing landrace diversity empowers wheat breeding." Nature 632, 823-831. doi: 10.1038/s41586-024-07682-9

Crossa, J., Perez-Rodriguez, P., Cuevas, J., Montesinos-Lopez, O., Jarquin, D., de los Campos, G., Burgueno, J., Gonzalez-Camacho, J. M., Perez-Elizalde, S., Beyene, Y., et al. (2017). "Genomic

535 selection in plant breeding: Methods, models, and perspectives." Trends in Plant Science 22, 961-
536 975. doi: 10.1016/j.tplants.2017.08.011

537 Crossa, J., Martini, J. W. R., Vitale, P., Perez-Rodriguez, P., Costa-Neto, G., Fritsche-Neto, R.,
538 Runcie, D., Cuevas, J., Toledo, F., Li, H., et al. (2025). "Expanding genomic prediction in plant
539 breeding: Harnessing big data, machine learning, and advanced software." Trends in Plant Science
540 30, 756-774. doi: 10.1016/j.tplants.2024.12.009

541 Fan, X., Liu, X., Feng, B., Zhou, Q., Deng, G., Long, H., Cao, J., Guo, S., Ji, G., Xu, Z., et al. (2022).
542 "Construction of a novel Wheat 55 K SNP array-derived genetic map and its utilization in QTL
543 mapping for grain yield and quality related traits." Frontiers in Genetics 13, 978880. doi:
544 10.3389/fgene.2022.978880

545 Gao, S., Yu, T., Rasheed, A., Wang, J., Crossa, J., Hearne, S., and Li, H. (2025). "Fast-forwarding
546 plant breeding with deep learning-based genomic prediction." Journal of Integrative Plant Biology
547 67, 1700-1705. doi: 10.1111/jipb.13914

548 Geurts, P., Ernst, D., and Wehenkel, L. (2006). "Extremely randomized trees." Machine Learning 63,
549 3-42. doi: 10.1007/s10994-006-6226-1

550 Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). "Genomic selection in plant
551 breeding: A comparison of models." Crop Science 52, 146-160. doi: 10.2135/cropsci2011.06.0297

552 International Wheat Genome Sequencing Consortium. (2018). "Shifting the limits in wheat research
553 and breeding using a fully annotated reference genome." Science 361, eaar7191. doi:
554 10.1126/science.aar7191

555 Jackson, R., Buntjer, J. B., Bentley, A. R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman,
556 E., Poupard, B., et al. (2023). "Phenomic and genomic prediction of yield on multiple locations in
557 winter wheat." Frontiers in Genetics 14, 1164935. doi: 10.3389/fgene.2023.1164935

558 Jiao, C., Xie, X., Hao, C., Chen, L., Xie, Y., Garg, V., Zhao, L., Wang, Z., Zhang, Y., Li, T., et al.
559 (2025). "Pan-genome bridges wheat structural variations with habitat and breeding." Nature 637,
560 384-393. doi: 10.1038/s41586-024-08277-0

561 Jubair, S., and Domaratzki, M. (2023). "Crop genomic selection with deep learning and
562 environmental data: A survey." Frontiers in Artificial Intelligence 5, 1040295. doi:
563 10.3389/frai.2022.1040295

564 Kaushal, S., Gill, H. S., Billah, M. M., Khan, S. N., Halder, J., Bernardo, A., Amand, P. S., Bai, G.,
565 Glover, K., Maimaitijiang, M., et al. (2024). "Enhancing the potential of phenomic and genomic
566 prediction in winter wheat breeding using high-throughput phenotyping and deep learning." Frontiers
567 in Plant Science 15, 1410249. doi: 10.3389/fpls.2024.1410249

568 Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017).
569 "LightGBM: A highly efficient gradient boosting decision tree." Advances in Neural Information
570 Processing Systems 30, 3146-3154.

571 Kick, D. R., Wallace, J. G., Schnable, J. C., Kolkman, J. M., Alaca, B., Beissinger, T. M., Edwards,
572 J., Ertl, D., Flint-Garcia, S., Gage, J. L., et al. (2023). "Yield prediction through integration of

genetic, environment, and management data through deep learning." G3: Genes, Genomes, Genetics 13, jkad006. doi: 10.1093/g3journal/jkad006

Liu, J., Gock, A., Ramm, K., Stops, S., Phongkham, T., Norman, A., Eastwood, R., Stone, E., and Dillon, S. (2025). "Incorporating gene expression and environment for genomic prediction in wheat." Frontiers in Plant Science 16, 1506434. doi: 10.3389/fpls.2025.1506434

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems 30, 4765-4774.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). "Prediction of total genetic value using genome-wide dense marker maps." Genetics 157, 1819-1829. doi: 10.1093/genetics/157.4.1819

Mo, Z., Zhu, J., Wei, J., Zhou, J., Xu, Q., Tang, H., Mu, Y., Deng, M., Jiang, Q., Liu, Y., et al. (2021). "The 55K SNP-based exploration of QTLs for spikelet number per spike in a tetraploid wheat population." Frontiers in Plant Science 12, 732837. doi: 10.3389/fpls.2021.732837

Montesinos-Lopez, A., Rivera, C., Pinto, F., Pinera, F., Gonzalez, D., Reynolds, M., Perez-Rodriguez, P., Li, H., Montesinos-Lopez, O. A., and Crossa, J. (2023). "Multimodal deep learning methods enhance genomic prediction of wheat breeding." G3: Genes, Genomes, Genetics 13, jkad045. doi: 10.1093/g3journal/jkad045

Montesinos-Lopez, A., Crespo-Herrera, L., Dreisigacker, S., Gerard, G., Vitale, P., Saint Pierre, C., Govindan, V., Tadesse, Z. T., Flores, M. C., Perez-Rodriguez, P., et al. (2024a). "Deep learning methods improve genomic prediction of wheat breeding." Frontiers in Plant Science 15, 1324090. doi: 10.3389/fpls.2024.1324090

Montesinos-Lopez, O. A., Montesinos-Lopez, A., Perez-Rodriguez, P., Barron-Lopez, J. A., Martini, J. W. R., Fajardo-Flores, S. B., Gaytan-Lugo, L. S., Santana-Mancilla, P. C., and Crossa, J. (2021). "A review of deep learning applications for genomic selection." BMC Genomics 22, 19. doi: 10.1186/s12864-020-07319-x

Montesinos-Lopez, O. A., Chavira-Flores, M., Kiasmiantini, Crespo-Herrera, L., Saint Pierre, C., Li, H., Fritsche-Neto, R., Al-Nowibet, K., Montesinos-Lopez, A., and Crossa, J. (2024b). "A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding." Genetics 228, iyae161. doi: 10.1093/genetics/iyae161

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12, 2825-2830.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nature Genetics 38, 904-909. doi: 10.1038/ng1847

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). "CatBoost: Unbiased boosting with categorical features." Advances in Neural Information Processing Systems 31, 6638-6648.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., and He, Z. (2017). "Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives." Molecular Plant 10, 1047-1064. doi: 10.1016/j.molp.2017.06.008

Rasheed, A., and Xia, X. (2019). "From markers to genome-based breeding in wheat." Theoretical and Applied Genetics 132, 767-784. doi: 10.1007/s00122-019-03286-4

Semagn, K., Iqbal, M., Alachiotis, N., N'Diaye, A., Pozniak, C., and Spaner, D. (2021). "Genetic diversity and selective sweeps in historical and modern Canadian spring wheat cultivars using the 90K SNP array." Scientific Reports 11, 23773. doi: 10.1038/s41598-021-02666-5

Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N., and Chen, F. (2020). "The Wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat." Plant Biotechnology Journal 18, 1354-1360. doi: 10.1111/pbi.13361

Togninalli, M., Wang, X., Kucera, T., Shrestha, S., Juliana, P., Mondal, S., Pinto, F., Govindan, V., Crespo-Herrera, L., Huerta-Espino, J., et al. (2023). "Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics." Bioinformatics 39, btad336. doi: 10.1093/bioinformatics/btad336

Vitale, P., Montesinos-Lopez, O., Gerard, G., Velu, G., Tadesse, Z., Montesinos-Lopez, A., Dreisigacker, S., Pacheco, A., Toledo, F., Saint Pierre, C., et al. (2025). "Improving wheat grain yield genomic prediction accuracy using historical data." G3: Genes, Genomes, Genetics 15, jkaf038. doi: 10.1093/g3journal/jkaf038

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., et al. (2020). "Multiple wheat genomes reveal global variation in modern breeding." Nature 588, 277-283. doi: 10.1038/s41586-020-2961-x

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., et al. (2014). "Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array." Plant Biotechnology Journal 12, 787-796. doi: 10.1111/pbi.12183

Xiong, H., Li, Y., Guo, H., Xie, Y., Zhao, L., Gu, J., Zhao, S., Ding, Y., and Liu, L. (2021). "Genetic mapping by integration of 55K SNP array and KASP markers reveals candidate genes for important agronomic traits in hexaploid wheat." Frontiers in Plant Science 12, 628478. doi: 10.3389/fpls.2021.628478

Yannam, V. R. R., Lopes, M. S., and Soriano, J. M. (2025). "Optimizing genomic selection models for wheat breeding under contrasting water regimes in a mediterranean environment." Plant Methods 21, 95. doi: 10.1186/s13007-025-01467-5

Zhou, C., Xiong, H., Fu, M., Guo, H., Zhao, L., Xie, Y., Gu, J., Zhao, S., Ding, Y., Li, Y., et al. (2023). "Genetic mapping and identification of Rht8-B1 that regulates plant height in wheat." BMC Plant Biology 23, 333. doi: 10.1186/s12870-023-04343-3

Figures

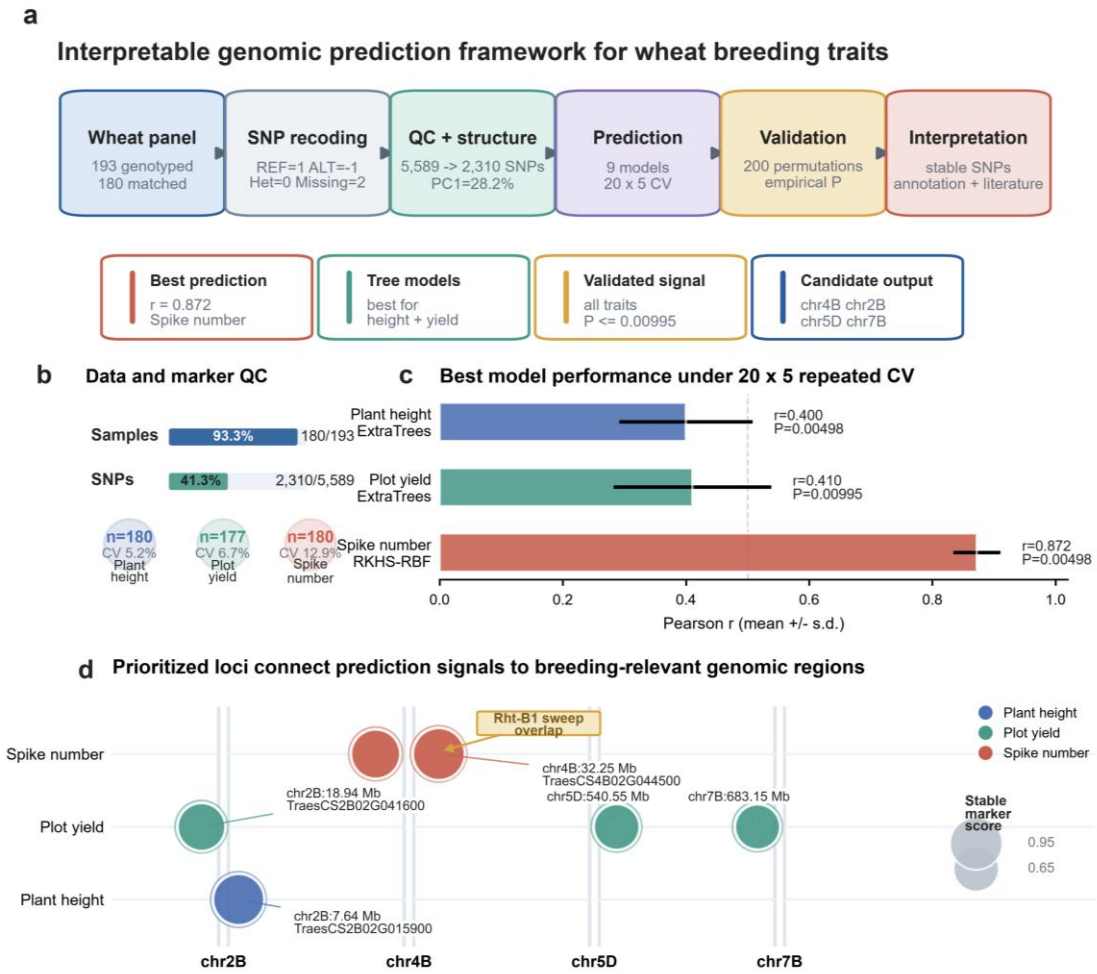
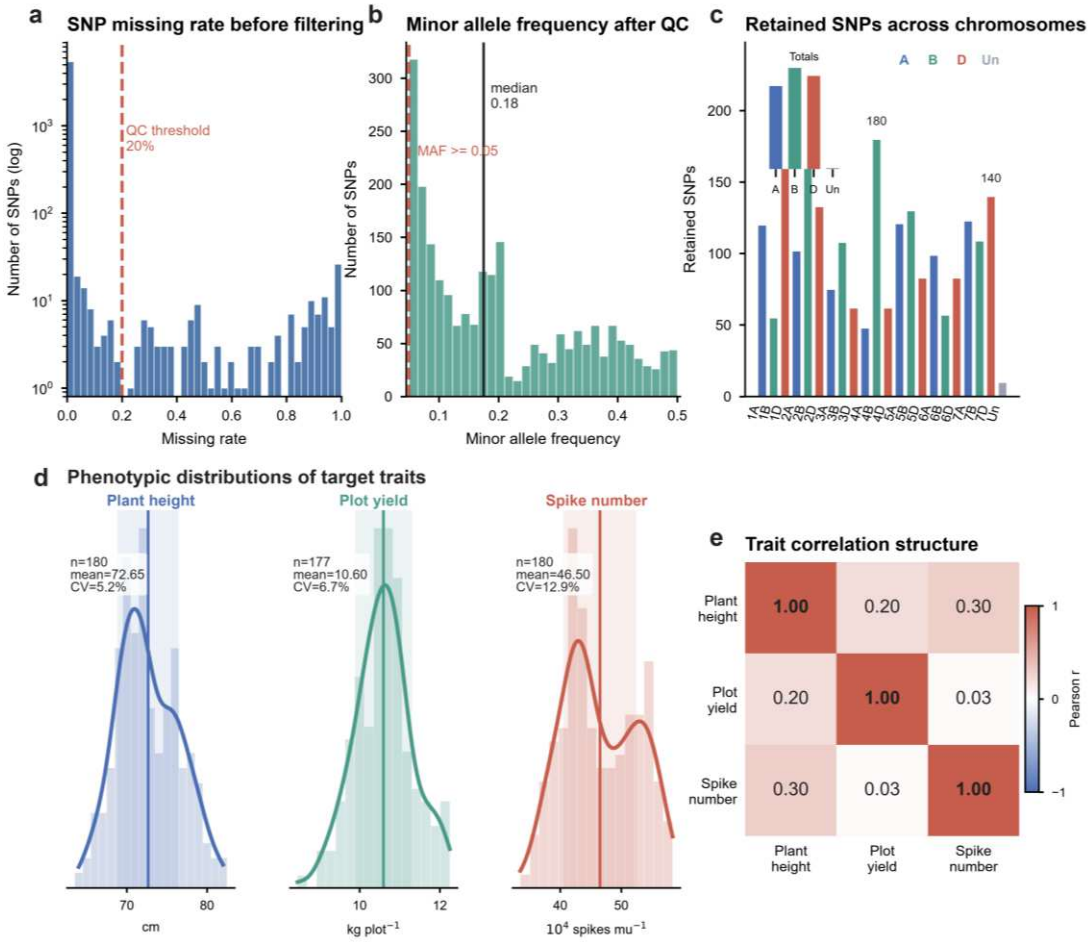


Figure 1. Workflow of SNP-based genomic prediction in wheat. (a) Study workflow, (b) dataset and SNP quality-control summary, (c) cross-validation performance, and (d) candidate locus overview. This workflow links preprocessing, validation and candidate-locus interpretation.



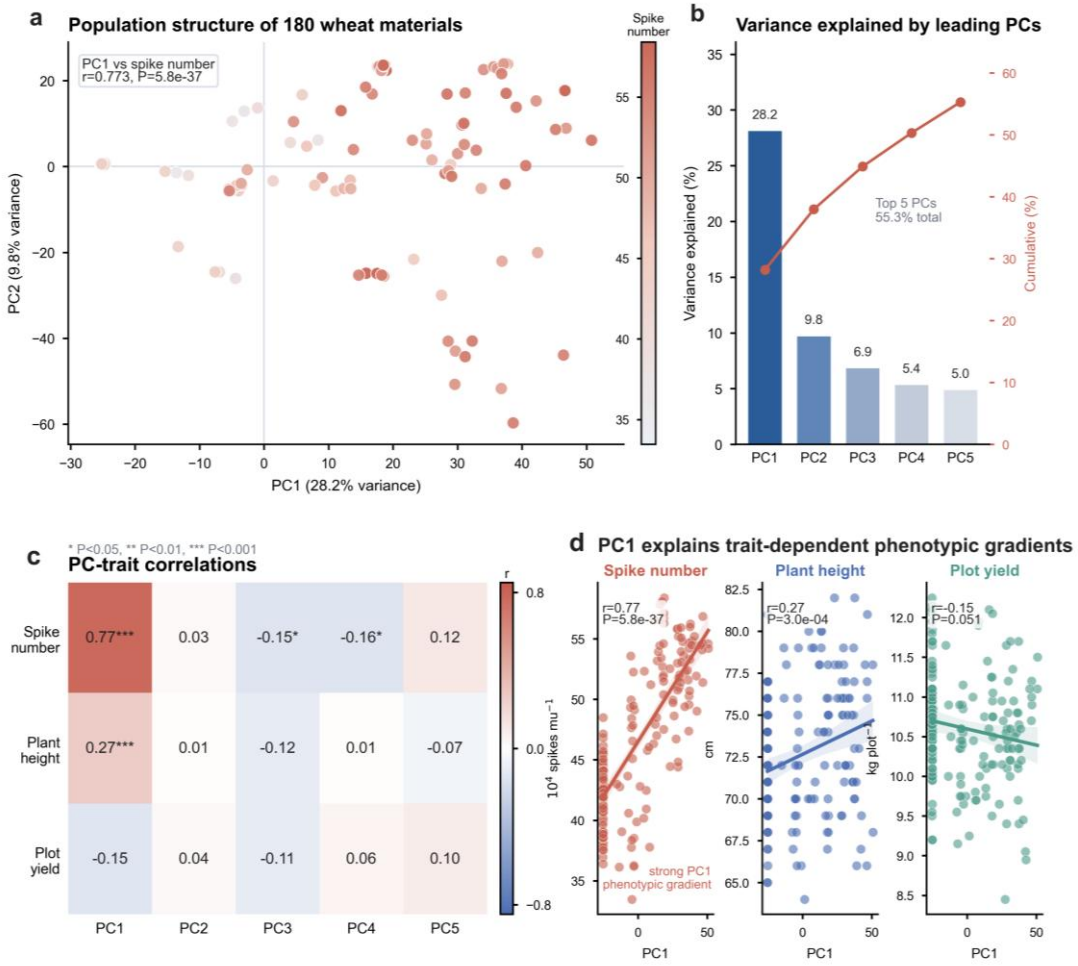
654

655 **Figure 2**. Genotype quality control and phenotype distribution. (a) SNP missing rate, (b) minor allele
656 frequency, (c) chromosome-wise SNP distribution, (d) trait distributions, and (e) trait correlation
657 matrix. This figure summarizes marker quality, genome coverage and trait variation.

658

659

660



661

662 **Figure 3**. Population structure of the wheat panel. (a) PCA of wheat materials, (b) variance explained
663 by leading PCs, (c) PC-trait correlations, and (d) PC1-trait relationships. This figure shows
664 population structure and trait associations.

665

666

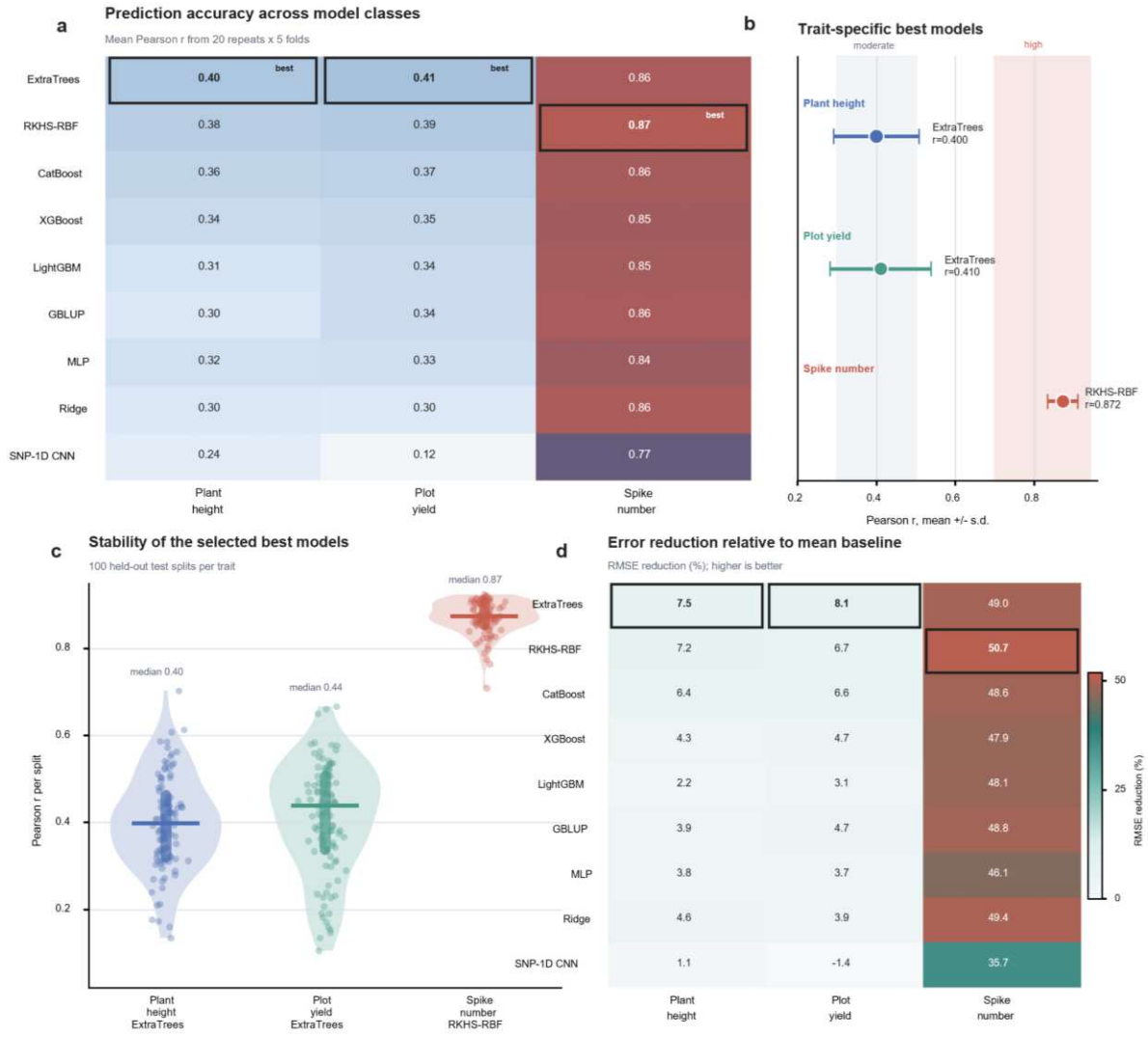


Figure 4. Genomic prediction performance across models. (a) Model-by-trait prediction accuracy, (b) selected model performance for each trait, (c) prediction stability across test splits, and (d) RMSE reduction relative to the mean baseline. This figure compares accuracy and error reduction across models.

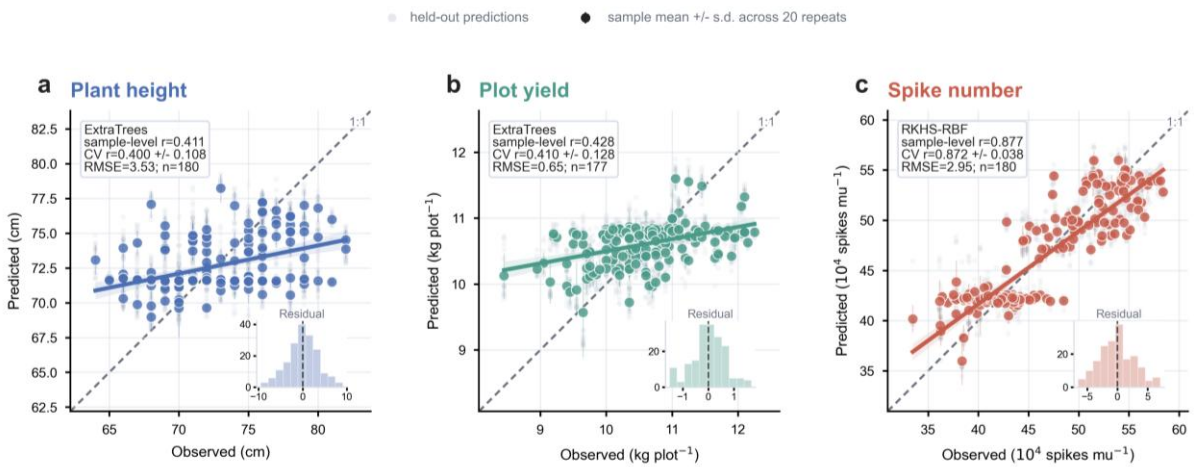


Figure 5. Observed and predicted values for the three traits. (a) Plant height, (b) plot yield, and (c) spike number. This figure shows agreement between observed and predicted trait values.

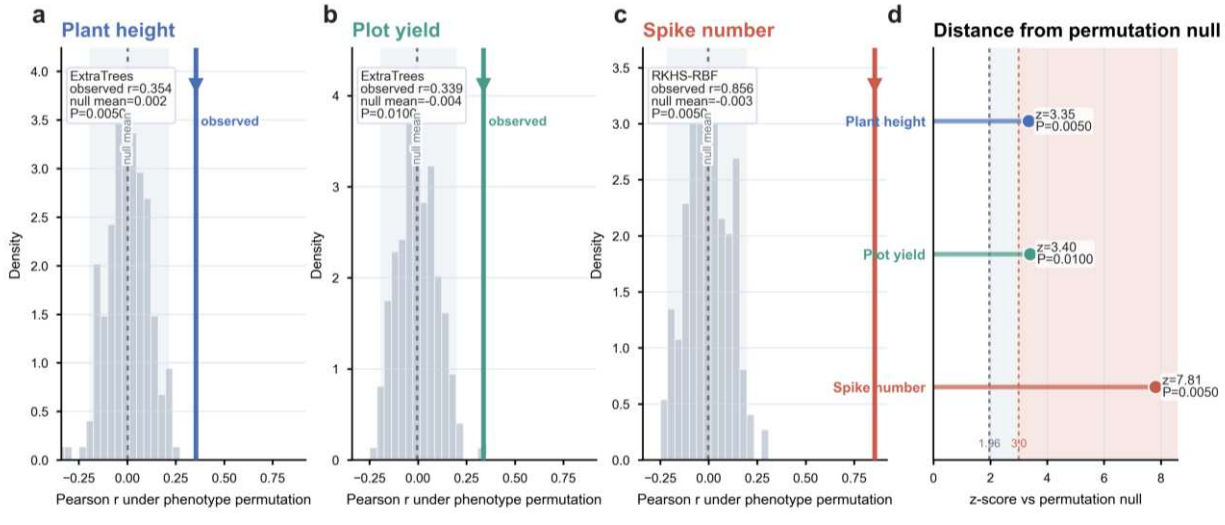


Figure 6. Permutation test of genomic prediction accuracy. (a) Plant height, (b) plot yield, (c) spike number, and (d) distance of observed accuracy from the permutation null distribution. This figure validates prediction accuracy against permutation null distributions.

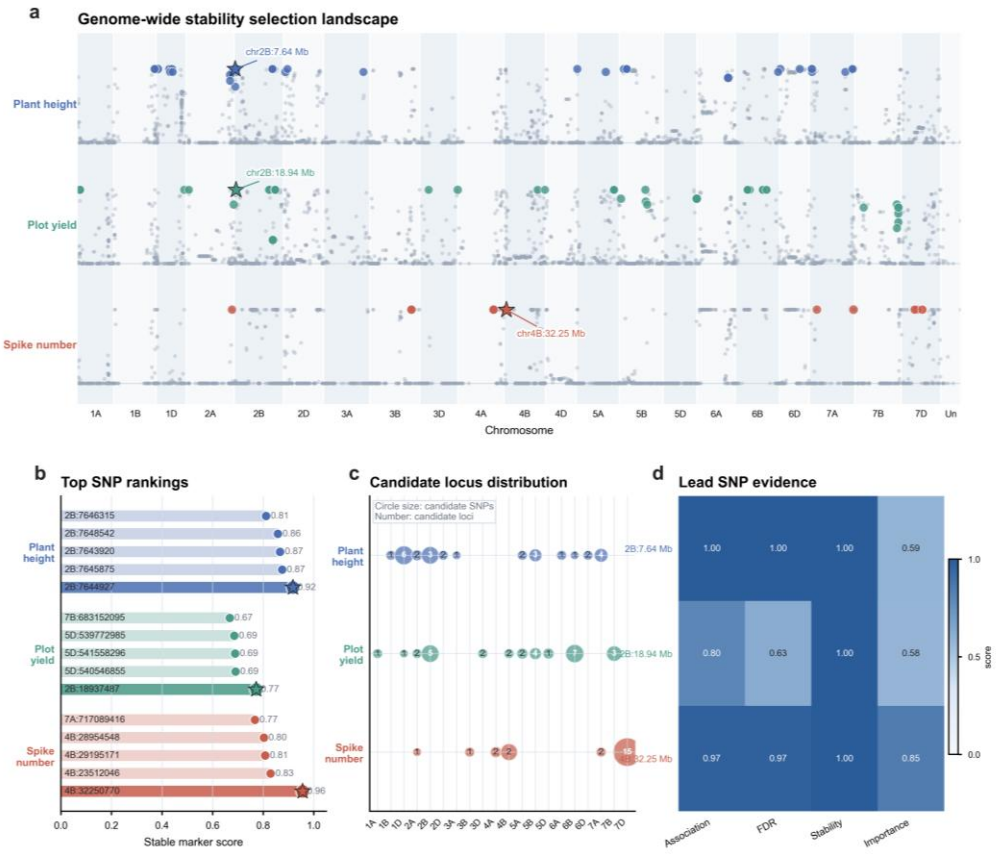


Figure 7. Candidate SNP prioritization for agronomic traits. (a) Genome-wide stability selection landscape, (b) top-ranked SNPs, (c) candidate locus distribution, and (d) lead SNP evidence profiles. This figure summarises evidence for genome-wide candidate SNP prioritisation.

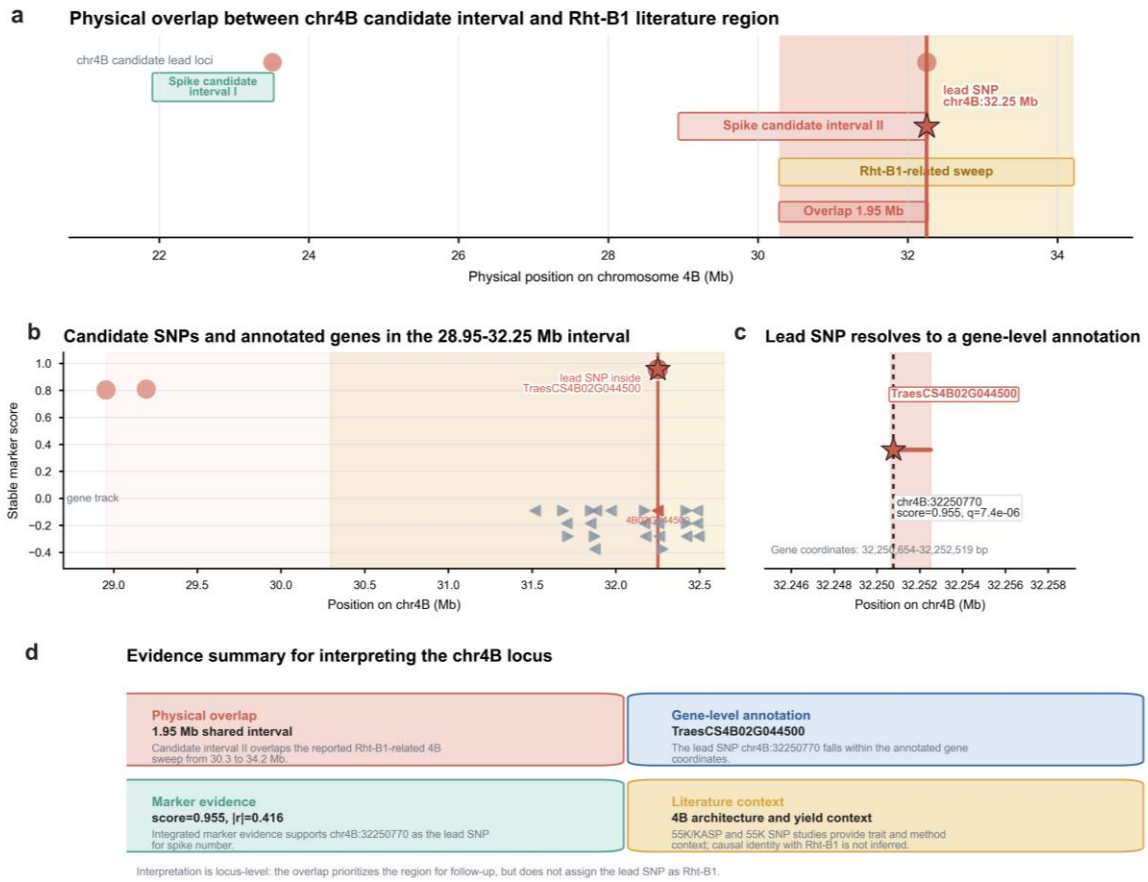


Figure 8. Annotation of the chr4B candidate region. (a) Overlap between the candidate interval and the Rht-B1-related region, (b) candidate SNPs and annotated genes, (c) gene-level annotation of chr4B:32250770, and (d) locus-level evidence summary. This figure contextualises the chr4B signal with annotation and literature evidence.

Tables

Table 1. Dataset composition, SNP quality control and phenotype summary. This table defines the analysed samples, retained markers and phenotype summaries.

Category	Item	n/count	Summary
Genotype data	Genotyped wheat materials	193	55K SNP-chip dataset
Phenotype data	Phenotyped materials	180	Matched agronomic records
Integrated dataset	Matched genotype-phenotype samples	180	Used for genomic prediction
SNP markers	Raw SNPs	5,589	Before quality control
SNP markers	Retained SNPs	2,310	After quality control
SNP markers	Removed SNPs	3,279	Excluded by QC filters
Trait	Plant height	180	72.65 ± 3.81 cm; range 64.00-82.00

Category	Item	n/count	Summary
Trait	Plot yield	177	10.60 ± 0.71 kg plot^{-1}; range 8.45-12.25
Trait	Spike number	180	46.50 ± 5.99 ×10^{4} spikes mu^{-1}; range 33.46-58.41

Table 2. Trait-specific genomic prediction performance and permutation-test significance. This table reports model selection, prediction accuracy and empirical significance.

Trait	n	Selected model	Pearson r	RMSE	MAE	Empirical P
Plant height	180	ExtraTrees, top 500 SNPs	0.400 ± 0.108	3.526 ± 0.399	2.759 ± 0.344	0.00498
Plot yield	177	ExtraTrees, top 500 SNPs	0.410 ± 0.128	0.652 ± 0.077	0.512 ± 0.065	0.00995
Spike number	180	RKHS-RBF, all QC SNPs	0.872 ± 0.038	2.954 ± 0.341	2.360 ± 0.289	0.00498

Table 3. Lead candidate SNPs, genomic annotation and literature context. This table links lead SNPs to candidate intervals, gene annotations and literature context.

Trait	Lead SNP	Candidate interval	Stable marker score	r	q value	Gene annotation	Literature context
Plant height	chr2B:7,644,927	chr2B:7,643,920-10,179,688	0.918	0.342	2.05 × 10^{-3}	Inside TraesCS2B02G015900	Trait/method-level support
Plot yield	chr2B:18,937,487	chr2B:18,937,487-21,650,737	0.773	0.309	3.07 × 10^{-3}	Nearest to TraesCS2B02G041600, 595 bp	Trait/method-level support
Spike number	chr4B:32,250,770	chr4B:28,954,548-32,250,770	0.955	0.416	7.36 × 10^{-6}	Inside TraesCS4B02G044500	Overlaps the Rht-B1-related 4B region