


Genomic selection accuracy for the yield trait in alfalfa half-sib families based on GWAS strategies

Yangyang Han^{1,2} , Bao Ao^{1,2}, Fan Zhang^{1,2}, Song Jing^{1,2}, Pan Xu^{1,2}, Qi Yan^{1,2}, Boyu Chu^{1,2} and Jiyou Zhang^{1,2*}

¹ State Key Laboratory of Grassland Agro-ecosystems, Lanzhou 730020, China

² College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou 730020, China

* Corresponding author, E-mail: zhangjiyou@lzu.edu.cn

Abstract

Alfalfa (*Medicago sativa* L.), a globally vital perennial forage supporting livestock, faces stagnant yield improvement due to polyploid complexity and slow genetic gains. To enhance genomic selection (GS) for fresh weight yield, this study integrated genome-wide association study (GWAS) strategies in 136 alfalfa half-sib families. Phenotypic data from five cuts collected during 2020–2022 were modeled into Best Linear Unbiased Predictions (BLUPs) to control environmental noise. GWAS, using these BLUPs, identified 19 significant single nucleotide polymorphisms (SNPs) associated with fresh weight per plant (FW). The prediction accuracy of 13 GS models (including gBLUP, rrBLUP, Bayesian, and machine learning methods) was evaluated. Strikingly, GWAS-informed SNP subsets boosted accuracy to 0.85–0.90. Moreover, the wgBLUP model, incorporating SNP-trait association strength as genomic weights, achieved high accuracy (0.90) even with all SNPs. These results demonstrate that leveraging GWAS data through SNP prioritization or weighted matrices significantly overcomes historical breeding bottlenecks and accelerates genetic gain, while this study proposes wgBLUP as an optimal strategy for accelerating alfalfa yield breeding in half-sib populations.

Citation: Han Y, Ao B, Zhang F, Jing S, Xu P, et al. 2025. Genomic selection accuracy for the yield trait in alfalfa half-sib families based on GWAS strategies. *Grass Research* 5: e030 <https://doi.org/10.48130/grares-0025-0028>

Introduction

The global population has exceeded eight billion and continues to rise, placing unprecedented pressure on food and forage systems worldwide. This growing demand underscores the urgent need to improve agricultural productivity and sustainability. Alfalfa (*Medicago sativa* L.) plays a crucial role in this context. As one of the most widely cultivated and economically important perennial leguminous forage crops, alfalfa is prized for its high biomass yield, nutritional quality, and ecological benefits, earning it the title of the 'Queen of Forages'^[1,2]. It is rich in crude protein, vitamins, and minerals, making it an ideal feed for dairy and meat livestock, and it is foundational to the global dairy industries^[3,4]. Additionally, alfalfa improves soil fertility through symbiotic nitrogen fixation, enhances water use efficiency, and supports agroecological resilience by reducing erosion and enhancing carbon sequestration^[5,6]. Despite its numerous agronomic and environmental advantages, the genetic improvement of alfalfa has been relatively slow compared to major cereal crops. This lag is mainly due to its complex genetic architecture and biological characteristics. As an autotetraploid and outcrossing species, alfalfa suffers from severe inbreeding depression and exhibits substantial non-additive genetic variance^[7]. The long breeding cycles, low heritability of key traits, and high levels of genotype-by-environment interaction further complicate conventional phenotypic selection, leading to stagnant or marginal yield gains^[3,7].

To address these challenges, genomic tools have been increasingly adopted to enhance the efficiency of alfalfa breeding programs. Among them, genomic selection (GS) has emerged as a powerful and promising method. First proposed by Meuwissen et al., GS enables the prediction of breeding values using genome-wide molecular markers without the need to identify specific trait-linked loci^[8]. Unlike traditional marker-assisted selection (MAS),

which focuses on a few significant markers with large effects, GS captures the collective contribution of thousands of genome-wide single nucleotide polymorphisms (SNPs), including those with small or moderate effects^[8,9]. This is especially beneficial for complex traits like biomass yield, stress tolerance, and persistence, which are controlled by many genes and are influenced by environmental variation^[9,10].

The successful implementation of GS relies on several factors, including population structure, training population size, marker density, statistical model choice, and trait heritability^[11,12]. In alfalfa, GS has been applied to full-sib and half-sib populations for yield and quality traits, demonstrating moderate to high prediction accuracies depending on the model used^[9,13]. Advanced statistical models such as genomic BLUP (gBLUP), ridge regression BLUP (rrBLUP), Bayesian regressions, and machine learning methods (e.g., support vector machines and random forest) have all been tested for alfalfa genomic prediction^[9,14]. Nazzicari et al. showed that modeling genome dosage using allele ratios in autotetraploid alfalfa improved predictive ability compared with simpler diploid parametrizations^[15]. Moreover, recent studies have emphasized the importance of genome parametrization, marker filtering, and SNP selection strategies to further enhance prediction accuracy.

While GS functions independently of trait-specific loci, incorporating information from genome-wide association studies (GWAS) can refine marker selection and enhance model performance. GWAS has been widely applied in alfalfa to detect SNPs associated with yield, leaf size, flowering time, fall dormancy, salt tolerance, and forage quality traits^[9,16,17]. For instance, the identification of trait-associated SNPs through GWAS allows researchers to build GS models that prioritized in GS models to emphasize informative genomic regions, potentially improving accuracy, especially when training populations are small or trait heritability is low. Medina et al. pioneered the integration of GWAS-derived *p*-values into GS models for alfalfa,

resulting in improved prediction accuracy for forage yield^[18]. Similarly, Zhang et al. demonstrated that using the top 3,000 trait-associated SNPs selected via GWAS combined with support vector machine regression yielded significantly improved genomic prediction for fall dormancy^[17]. Although GWAS provides insight into the genetic architecture of complex traits, it often lacks the power to detect loci of small effect and may not capture the full polygenic basis of quantitative traits. Therefore, GS remains the preferred approach for complex trait prediction in breeding applications, and GWAS can serve as a complementary tool to optimize SNP selection. The synergy between these methods is especially valuable in polyploid crops like alfalfa, where complex inheritance patterns and high genetic variability challenge traditional selection strategies^[7,15,17].

This study investigates the impact of different prediction models and strategies on GS prediction accuracy, particularly focusing on the effect of selecting varying numbers of SNPs derived from GWAS. Based on a half-sib population of 136 families, the study aims to explore how these strategies influence the efficiency and accuracy of GS, with a focus on improving yield traits in alfalfa. Additionally, the potential of integrating GWAS-derived SNP subsets into GS models is assessed, evaluating their impact on genomic prediction accuracy and providing insights for future breeding programs.

Materials and methods

Plant materials and field experiment

An alfalfa association panel of germplasms, consisting of 136 half-sib families selected from the Italy Centre for Fodder Crops and Dairy Productions, was used in this experiment. They are F2 generation plants of ten elite cultivars from different European countries, each representing or containing superior genetic resources from a specific region^[19]. The development of this panel began with an initial round of random cross-pollination among these ten cultivars, which produced the F1 generation. From this F1 population, 136 superior individual plants were selected as mother plants. These selected mother plants were then subjected to open pollination (mixed crossing) in a common field. Seeds were harvested from each mother plant individually, thereby forming the 136 distinct half-sib families that constituted the plant materials for this study.

The panel of germplasms was planted in Yuzhong County, Gansu Province. A completely randomized design was used in the field trial. The experiment had three replications, and all germplasm had a row containing 20 plants in each replication. The spacing between plants was 15 and 30 cm between rows. The annual rainfall of Yuzhong (longitude 104.09° E, latitude 35.87° N) was 380 mm. Four check alfalfa cultivars were also randomly planted in each replication, namely 'Gongnong No.1', 'LM809', 'Hezuo', and 'Hulun Buir'.

Phenotyping and data analysis

Whenever alfalfa reached its early flowering stage, three plants were randomly selected from each row of each replication to measure the yield traits after removing three plants from the head and tail. The fresh weight per plant (FW) for one cut in 2020 (the first cut) and two cuts (the first and second cuts) in 2021 and 2022 were measured. DeltaGen (<https://deltagen.agr.nz/app/deltagen>) was used for the variance analysis of phenotypic data, and the BLUPs were calculated. The check cultivars were assumed to be a fixed effect, and other factors, such as genotype, year, and replication, were treated as random effects. The following random-effects model was utilized for BLUP:

$$y_{ijhln} = \mu + g_i + e_j + s_{hl} + g_i \times e_j + r_n + \varepsilon_{ijhln}$$

where, y_{ijhln} represents the phenotype of the i th genotype in the n -th replication located in h th row and l th column in the j -th year, μ represents the mean value after considering the fixed effect of the check cultivars, g_i represents the genetic effect of the i -th genotype, e_j represents the effect of the j -th year, s_{hl} represents the soil zone effect of the h -th row and l -th column, $g_i \times e_j$ represents the interaction effect of the i -th genotype and the j -th year, r_n represents the block effect of the n -th replication, and ε_{ijhln} is the random error.

Based on the division of phenotypic variance components, broad-sense heritability (H^2) was calculated using σ_g^2 , σ_{gy}^2 , σ_e^2 , and the formula is as follows:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{gy}^2}{n_y} + \frac{\sigma_e^2}{n_y n_r}}$$

where, σ_g^2 , σ_{gy}^2 , and σ_e^2 , are the variance components of genotype, genotype by year, and residual error, respectively; n_y is the number of years and n_r is the number of blocks (replications).

Then, the initial phenotypic data and BLUPs were used to analyze the phenotypic and genotypic correlations of traits in different environments, using Pearson coefficients. Cluster analysis of the 136 germplasms was performed using a heat map and principal component analysis (PCA) based on BLUPs.

SNP calling and filtering

DNA extraction, library preparation, and sequencing have been completed by Paolo's team from the Italy Centre for Fodder Crops and Dairy Productions, and the obtained sequencing data have been deposited in the NCBI RSA archive under submission number PRJNA1092606^[15]. Nucleotide variants were identified using the UnifiedGenotyper tool within the Genome Analysis Toolkit (GATK v3.4-46)^[20] with the following parameter settings: -stand_call_conf 50.0, -stand_emit_conf 10.0, -dcov 1000, -A Coverage, and -A Allele-Balance. Raw SNP variants were filtered using the GATK VariantFiltration tool with the following criteria: QD < 2.0, FS > 60.0, SOR > 3.0, MQ < 40.0, and MQRankSum < -12.5. A total of 1,360,956 SNPs were retained after this step. As the reference genome, the sequence obtained from cultivated alfalfa (cultivar XinJiangDaYe)^[21], which consists of the full resequencing of all four copies of each chromosome, was used. For read alignment, the longest copy for each chromosome was selected. The obtained single nucleotide polymorphisms (SNPs) were filtered based on minor allele frequency (MAF ≥ 10%), missing data per marker (≤ 10%), and the elimination of third-order loci. After filtration, a total of 211,610 high-quality SNPs were obtained from 136 germplasms, and they will be further analyzed.

Population structure analysis and GWAS

The genotypic data were used for population structure analysis using the ADMIXTURE software (version 1.3.0). The variant call format (VCF) file was first converted to PLINK binary format using PLINK (version 1.90b7.1) with the following parameters: --vcf input.vcf --make-bed --out output --allow-extra-chr. To reduce linkage disequilibrium effects, the dataset was pruned using an iterative pruning approach with parameters --indep-pairwise 50 10 0.1. Population structure was inferred by running ADMIXTURE with cross-validation (--cv flag) for K values ranging from one to five. The optimal number of ancestral populations (K value) was determined by identifying the K with the lowest cross-validation error.

The R package GAPIT was used for GWAS. Three models, including generalized linear model (GLM), mixed linear model (MLM), and Fixed and Random Model Circulating Probability Unification (Farm-CPU), were used for analysis of marker-trait associations.

Genomic selection models

The 13 prediction models used in the GS include: (1) genomic BLUP (gBLUP)^[22], which is implemented through the R language package 'GAPIT'. (2) ridge regression BLUP (rrBLUP)^[23], this model is implemented through the R language package 'rrBLUP'. (3) the Bayesian whole-genome regression: Bayes A^[8], Bayes B^[24], Bayes C^[24], Bayesian Ridge Regression (BRR)^[25], Bayes LASSO (BL)^[26], and the bayes model is implemented through the R pack 'BGLR'^[27] (4) machine learning models^[17]: Ridge regression (Ridge), kernel ridge regression (KernelRidge), partial least squares regression (PLSR-regression), linear support vector regression (SVR_linear), polynomial kernel support vector regression (SVR_poly), and linear regression (linear). These models are implemented using the Python package 'sklearn'. (5) weighted genomic BLUP (wgBLUP)^[28], this model reconstructs the G matrix in the gBLUP model with the $-\log_{10}(p\text{-value})$ of SNP as the weight:

$$G_w = \frac{(ZW)(ZW)^T}{\sum_i w_i^2 \cdot 2p_i(1-p_i)}$$

where, Z is the centralized genotype matrix, w_i is the weight of the i th SNP, W is the diagonal matrix composed of the weight vectors, T represents the transpose of the matrix, and p_i represents the allele frequency of the i -th SNP.

Cross-validation and SNP setting

A repeated 5-fold cross-validation (CV) scheme was used to evaluate the performance of the model. The folds were created using the createFolds function from the R caret package. Individuals were randomly partitioned into training and test sets without family stratification, as the population genetic analysis using ADMIXTURE revealed no significant substructure (the cross-validation error was minimized at $K = 1$). In each round of 5-fold CV, 80% of the 136 samples were randomly selected as the training set and the remaining 20% as the test set. Then, the training set was used to fit the 13 prediction models to estimate the SNP genetic effects on the phenotype. These estimated genetic effects were used to obtain the genetic estimated breeding values (GEBV) for the test set. Then, the GEBV in the test set was compared with the actual phenotype, and a Pearson's correlation coefficient (r) was calculated. This coefficient is an indicator of prediction accuracy. The prediction accuracy was thus evaluated at the individual plant level. The entire 5-fold CV process was repeated ten times. Furthermore, to verify the influence of SNP subsets on the prediction model, based on the p -values

of each SNP obtained from GWAS (MLM model), eight subsets of SNPs were set up for the construction of the GS model. They were all SNPs (211,610 SNPs, SNPset1), SNPs with $-\log_{10}(p) \geq 4$ (SNPset2), SNPs with $-\log_{10}(p) \geq 3$ (SNPset3), and top 300 (SNPset4), top 500 (SNPset5), top 1,000 (SNPset6), top 5,000 (SNPset7), and top 10,000 (SNPset8) association SNPs. Additionally, to compare the impact of different GWAS models on GS, SNP subsets were selected in the same manner based on the results from the FarmCPU and GLM models. As a comparison, corresponding numbers of randomly chosen SNPs were also selected for the 5-fold cross-validation GS analysis. The numbers of randomly selected SNPs were set to 15, 150, 300, 500, 1,000, 5,000, and 10,000, respectively. To ensure randomness, this random selection process was repeated ten times.

To explore the influence of the proportion size of the training population on the prediction accuracy of the models, 3-fold CV and 10-fold CV were also conducted. In the 3-fold CV, 136 half-sib families were evenly divided into three parts. Each time, one of them was selected as the test set, and the other two as the training set for the GS. This was conducted three times in total to ensure that each family participated in both the training and test groups. In the 10-fold cross-validation, the data was divided into ten parts and underwent ten GS operations. The CV for each model was repeated more than ten times.

Results

Phenotypic analysis

From 2020 to 2022, a total of five cuts of FW yields were recorded. It was found that the original data for each cut and year are similar to a normal distribution (Fig. 1a), and there were significant differences in the FW trait among the different cuts (Table 1). In 2020, only one cut was harvested, and its yield was the lowest. In 2021 and 2022, the FW yield of the first cut was much higher than that of the second cut. In addition, the FW yield of the regenerated alfalfa was similar between the same cut, indicating that the yield of regenerated alfalfa tends to stabilize. The standard error range of the yield phenotype of alfalfa under different environments was 1.74–4.14, with an average value of 2.83. The coefficient of variation ranged from 25.62% to 51.89%, with an average value of 33.30% (Table 1), indicating that there is significant variation in this trait among half-sib families.

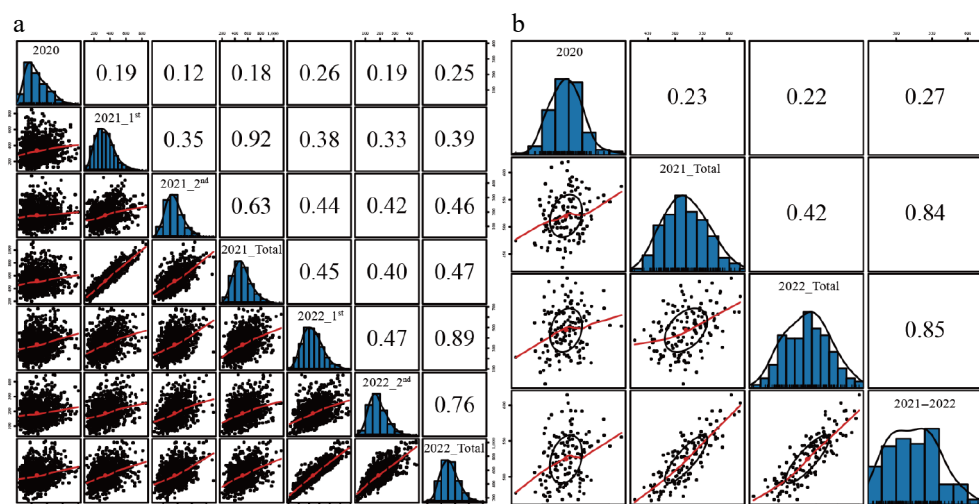


Fig. 1 The correlation of FW traits under different environments. (a) Using the original data. (b) Using the BLUP data. Note: The definition of the environment abbreviations (e.g., 2020, 2021_1st, 2022_Total) is identical to those provided in Table 1.

Table 1. The genotypic variance component of the FW trait of *M. sativa* half-sib families at Yuzhong in 2020–2022.

Trait	2020_FW	2021_1 st _FW	2021_2 nd _FW	2021_Total_FW	2022_1 st _FW	2022_2 nd _FW	2022_Total_FW	2021–2022_FW
Max	400	850	450	1,120	684	447	1,019	1,120
Min	20	110	40	200	112	46	165	165
Ave.	133.97	339.49	178.66	517.96	335.86	188.14	523.39	520.68
SE	1.98	3.46	1.74	4.14	2.83	1.85	3.82	2.82
CV (%)	51.89	35.73	34.43	27.96	29.59	34.36	25.62	26.8
σ^2_g	733.50 ± 115.76***	2,028.52 ± 417.90***	481.15 ± 106.55***	2,731.36 ± 582.07***	1,825.18 ± 329.27***	856.31 ± 148.04***	3,689.17 ± 630.95***	2,002.36 ± 456.00***
σ^2_{gy}	–	–	–	–	–	–	–	1211.48 ± 359.99**
σ^2_e	1,353.99 ± 61.52	10,076.00 ± 453.81	2,789.18 ± 125.55	14,255.73 ± 643.20	6,305.04 ± 283.56	2,395.75 ± 108.47	10,448.46 ± 470.62	12,473.92 ± 398.39
H ²	0.8298	0.6444	0.6082	0.6329	0.7226	0.7629	0.7606	0.6065

SE is the standard error, CV is the coefficient of variation, σ^2_g is the genotypic variance; σ^2_{gy} is the genotypic × year interaction variance; σ^2_e is the environmental variance; H^2 is the broad-sense heritability; 2020_FW, 2021_1st_FW, 2021_2nd_FW, 2021_Total_FW, 2022_1st_FW, 2022_2nd_FW, and 2021–2022_FW represent the FW data from the first cut in 2020, the first cut in 2021, the second cut in 2021, the sum of the two cuts in 2021, the first cut in 2022, the second cut in 2022, the sum of the two cuts in 2022, and the integrated value for 2021–2022, respectively. ** shows there were significant variations among 136 half-sib families (*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$).

Through the analysis of variance, it can be seen that the variation mainly existed among genotypes and the environment. There are significant differences in the genotypes of FW yields in each cut and the comprehensive environment ($p < 0.05$) (Table 1). Moreover, in the overall environment, this phenotype also shows significant differences in the interaction between genotype and year. After conducting the analysis of variance, the BLUP values of FW traits in the individual and comprehensive environments were calculated. The BLUP values of the traits were also close to the normal distribution (Fig. 1b, Supplementary Fig. S1).

The broad-sense heritability (H^2) was calculated based on the variance components. When disregarding the influence of cut and year, the H^2 of the five cuts ranges from 0.6 to 0.83, showing a relatively high heritability. After considering the influence of the year, its heritability is also at a relatively high level ($H^2 > 0.6$). These results suggest that phenotypic variation in FW yield is primarily due to genetic variation.

The FW yield was evaluated by employing BLUP. Correlation analysis was performed between different cuts and years of the half-sib population using the original data and BLUP values of the annual yield (Fig. 1). The correlation of different cuts and total yield of each year ranged from 0.12 to 0.89, indicating that there was a significant difference in the yield among different cuts and years. However, the yield correlation among the four cuts in 2021 and 2022 was higher than that in 2020 (Fig. 1a), indicating that the regenerative growth capacity of alfalfa cannot be equated with the initial growth capacity of the first cut. The yield correlation between BLUP values also validates this point, as demonstrated by the higher correlation between the annual yield and the combined 2021–2022 BLUPs compared to that with the first cut in 2020 (Fig. 1b). After including the correlations among the individual cut BLUP values in the analysis, the yield correlation coefficient between cuts did not increase and was lower than the correlation with the annual yield BLUPs.

Population structure analysis using SNPs

A total of 1.36 million SNPs were identified, from which 211,610 high-quality SNPs were retained. Chromosome 6 contained the highest number of SNPs, whereas chromosomes 3, 7, and 1 had relatively fewer (Fig. 2a). Among these, 164,862 SNPs were closely spaced (< 600 bp), with an average density of 34.93 bp/SNP. Notably, 139,467 SNPs were located within 50 bp of each other, with a mean density of 13.80 bp/SNP. In contrast, 46,740 SNPs were spaced over 600 bp apart, with an average interval of 15,167 bp (Supplementary Fig. S2). The minor allele frequency (MAF) distribution peaked at $MAF = 0.12$, beyond which the number of SNPs declined steadily with increasing MAF (Supplementary Fig. S2).

Based on the SNPs, the population structure of these 136 germplasms was analyzed. The results show that the cross-validation error is the lowest when the K value is 1 (Fig. 2b), indicating that all germplasms belong to a single genetic cluster without evidence of population substructure. This fits the characteristics of half-sib families ($K = 1$) (Fig. 2c). The results of population structure analysis showed that the population was a typical artificial genetic population after two generations of hybridization.

Association studies identify SNPs with different thresholds

Combining phenotypic BLUP values with the SNP data, GWAS was successfully performed using different models, and the p -value of 10^{-5} was used as the threshold for the selection of significant associations (Supplementary Figs S3–S5). In total, 19 significant SNPs were obtained that are associated with the FW traits of different environments, except for the FW traits of the second cut in 2021 (Table 2, Supplementary Table S1). Based on the GWAS results, several significant SNPs were located within annotated gene regions. The SNP Chr3.4_7976440 was identified in the promoter region of *MS.gene046331*, annotated as an LRR-repeat F-box protein. A cluster of SNPs on chromosome 2.4, such as Chr2.4_57419245, Chr2.4_57419282, and so on, was found within intronic regions of *MS.gene67840*, encoding an ankyrin repeat-containing protein. Additionally, Chr3.4_46433515 was located in an exon of the tryptophan synthase gene *MS.gene027196*, while Chr8.1_83204180 and Chr8.1_83204197 were both situated in the promoter region of the casein kinase gene *MS.gene072249*. The remaining significant SNPs were distributed in intergenic regions.

However, in the GLM and MLM, only for the first cut in 2021, the SNP related to the FW trait was identified. Among the three models, the FarmCPU identified the largest number of SNPs, while the MLM identified the least. As the threshold range expands, the number of SNPs will also increase. When the threshold range was $p < 10^{-4}$, $p < 10^{-3}$, and $p < 0.05$, the average number of SNPs provided by the FarmCPU model was 21.75,

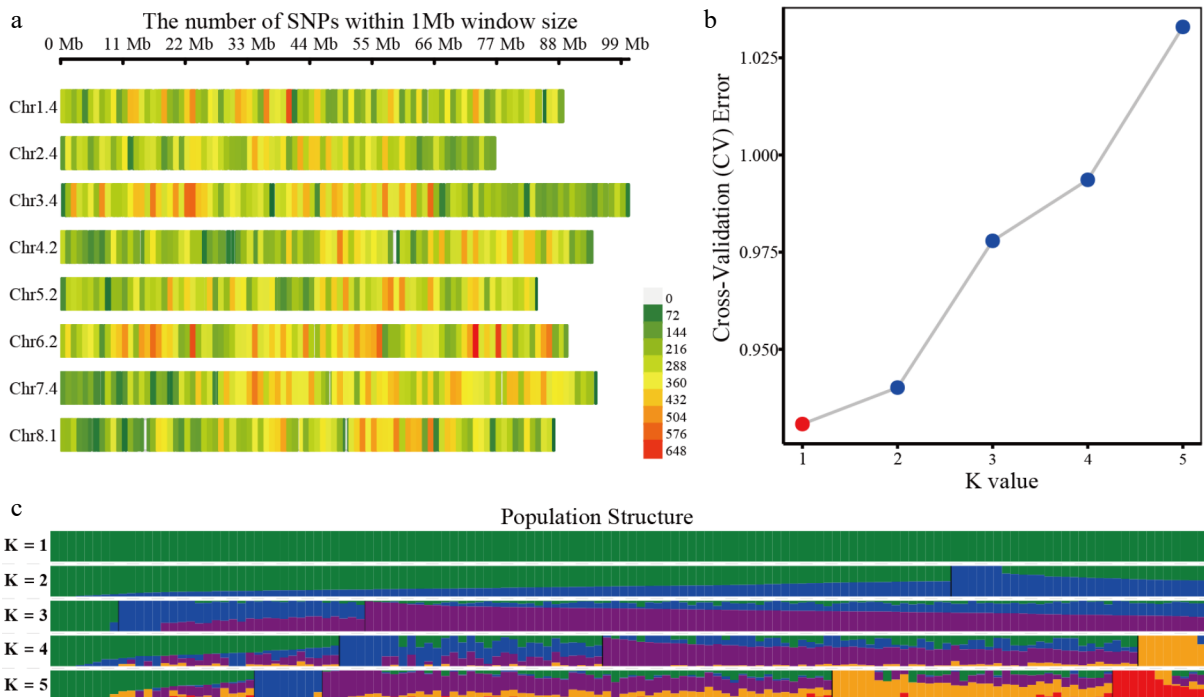


Fig. 2 Distribution of SNPs and population structure analysis of the genotypes of 136 alfalfa germplasms. (a) Distribution of SNPs on eight chromosomes. (b) Cross-validation analysis of the SNPs of half-sib families of *M. sativa*. (c) Population structure analysis using SNPs of half-sib families of *M. sativa*.

Table 2. The numbers of SNP associated with yield traits by using different thresholds.

Traits	SNP number											
	FarmCPU				GLM				MLM			
	10 ⁻⁵	10 ⁻⁴	10 ⁻³	0.05	10 ⁻⁵	10 ⁻⁴	10 ⁻³	0.05	10 ⁻⁵	10 ⁻⁴	10 ⁻³	0.05
2020_FW	2	23	251	10,554	0	10	179	10,518	0	4	162	9,973
2021_1 st _FW	7	17	214	10,563	7	10	143	10,295	6	11	132	9,884
2021_2 nd _FW	0	16	221	10,561	0	8	152	10,420	0	6	137	9,991
2021_Total_FW	3	27	221	10,543	0	16	146	10,276	0	16	140	10,137
2022_1 st _FW	2	21	206	10,562	0	13	113	9,836	0	13	113	9,836
2022_2 nd _FW	1	24	197	10,559	0	11	144	10,173	0	11	144	10,173
2022_Total_FW	3	24	182	10,549	0	14	119	10,222	0	13	120	10,161
2021–2022_FW	3	22	214	10,558	0	14	139	10,568	0	9	126	10,100

The definitions of abbreviations for each environmental trait (such as 2020_FW, 2021_1st_FW, 2022_Total_FW) are the same as those provided in Table 1. 10⁻⁵, 10⁻⁴, 10⁻³, and 0.05, represent that the set thresholds for *p*-value are 0.00001, 0.0001, 0.001, and 0.05 respectively.

213.25, and 10,556.13, respectively, while the GLM model was 12, 141.875, and 10,288.5, and the MLM model was 10.375, 134.25, and 10,031.88. This provides a basis for us when selecting the SNP subsets.

Genomic selection using all SNPs

Seven traditional models were used to conduct GS for the yield BLUP values in each cut, year, and integrated combined environment, including gBLUP, rrBLUP, and five Bayes models (Table 3). The prediction accuracy of GS across different environments ranged from 0.10 to 0.28, with mean accuracy ranging from 0.15 to 0.17 across models. Among all environments, the first cut in 2020 showed the highest prediction accuracy. Most models achieved accuracies between 0.24 and 0.28 in this environment, with the exception of Bayes Lasso.

Six machine learning models were applied for genomic prediction. Consistent with traditional methods, prediction accuracy across environments ranged from 0.12 to 0.28 (Table 4). The mean prediction accuracy of the models varied between 0.16 and 0.18.

The first cut in 2020 again showed the highest accuracy among all environments, with accuracies ranging from 0.25 to 0.28.

Genomic selection using the top SNP from GWAS

To improve the accuracy of GS prediction, different numbers of top-associated SNPs based on the GWAS results were selected, forming seven SNP subsets, from SNPset2 to SNPset8. These SNP subsets were used as genotype data to construct GS and conduct 5-fold cross-validation. To evaluate the potential impact of different GWAS models on subsequent GS accuracy, an analysis was conducted by constructing these SNP subsets based on the results of the FarmCPU, GLM, and MLM models, followed by gBLUP modeling. The results showed that when SNPs were selected using identical significance thresholds (SNPset2, SNPset3), the SNP sets identified by the FarmCPU model generally yielded the highest prediction accuracy, which was primarily attributable to its detection of a larger number of significant SNPs. However, when the number of selected SNPs was fixed across models (SNPset4–SNPset8), the prediction accuracies of the SNP subsets derived from

Table 3. The prediction accuracy of different GS models in different cuts and years.

Models	2020_FW	2021_1 st _FW	2021_2 nd _FW	2021_Total_FW	2022_1 st _FW	2022_2 nd _FW	2022_Total_FW	2021–2022_FW	Ave.
gBLUP	0.28	0.13	0.11	0.14	0.19	0.14	0.16	0.14	0.16
rrBLUP	0.27	0.14	0.14	0.14	0.17	0.10	0.11	0.13	0.15
BRR	0.27	0.15	0.13	0.13	0.16	0.13	0.12	0.14	0.16
BL	0.19	0.16	0.18	0.17	0.15	0.15	0.14	0.15	0.16
BayesA	0.26	0.16	0.18	0.15	0.17	0.13	0.12	0.15	0.17
BayesB	0.27	0.15	0.16	0.13	0.19	0.10	0.12	0.15	0.16
BayesC	0.27	0.12	0.13	0.15	0.18	0.12	0.12	0.16	0.16

The definitions of abbreviations for each environmental trait (such as 2020_FW, 2021_1st_FW, 2022_Total_FW) are the same as those provided in Table 1.

Table 4. The prediction accuracy of different machine learning models in different environments.

Models	2020_FW	2021_1 st _FW	2021_2 nd _FW	2021_Total_FW	2022_1 st _FW	2022_2 nd _FW	2022_Total_FW	2021–2022_FW	Ave.
Ridge	0.28	0.16	0.15	0.15	0.16	0.12	0.16	0.11	0.16
Kernel ridge	0.26	0.20	0.18	0.17	0.18	0.13	0.17	0.14	0.18
PLS regression	0.25	0.15	0.15	0.15	0.17	0.13	0.15	0.13	0.16
SVR_linear	0.28	0.15	0.14	0.19	0.19	0.14	0.13	0.12	0.17
SVR_poly	0.25	0.16	0.14	0.16	0.17	0.12	0.17	0.14	0.16
Linear	0.27	0.15	0.15	0.15	0.16	0.13	0.16	0.14	0.16

The definitions of abbreviations for each environmental trait (such as 2020_FW, 2021_1st_FW, 2022_Total_FW) are the same as those provided in Table 1.

the three GWAS models were highly comparable (Supplemental Table S1). Furthermore, a high degree of overlap (> 85%) was observed among the top 100, 500, and 1,000 SNP lists identified by the three models. This indicates that, provided the number of SNPs is controlled, the predictive performance of GS is largely insensitive to the choice of GWAS model. Given its conservative nature in controlling for false positives, the MLM model was therefore selected as the basis for all subsequent SNP filtering and GS analysis.

Subsequently, all 13 prediction models were used to perform GS utilizing these SNP subsets derived from the MLM model. Compared to using all SNPs (SNPset1), the GS models using SNP subsets showed a marked improvement in prediction accuracy (Fig. 3). Moreover, when the number of sets is small, as the number of SNPs in the subsets increases, the prediction accuracy shows a tendency to rise and then stabilizes. When the number of top SNPs selected by the seven models was greater than 1,000 but less than 10,000, their prediction accuracy was relatively stable, with the average values ranging from 0.85 to 0.9. For the YZ_2021–2022_FW trait specifically, prediction accuracy increased by approximately 0.70 compared to models using all SNPs (Fig. 3).

Similarly, six machine learning models were evaluated using the same SNP subsets. Compared to SNPset1, all machine learning models showed significantly improved prediction accuracy when using the SNP subsets, with the exception of KernelRidge, which exhibited lower accuracy when the number of SNPs in the subset was small. However, as the number of top SNPs increased beyond a certain threshold, the prediction accuracy of KernelRidge improved substantially, eventually reaching a level close to that of the other machine learning models (Fig. 4). The average prediction accuracy of the machine learning models from SNPset2 to SNPset8 in all environments was 0.42, 0.58, 0.69, 0.75, 0.79, 0.83, 0.85, and 0.85, respectively. Compared to all SNPs, the accuracy increased by up to 0.69.

To investigate whether other SNPs could yield similar effects, SNPs were randomly selected for GS on the YZ_2020_FW trait. Corresponding to SNPset2 to SNPset8, comparable numbers of SNPs were randomly chosen. After ten rounds of random selection and the same 5-fold cross-validation, the results indicated that randomly selected SNPs did not enhance prediction accuracy (Supplementary Table S2). On the contrary, when the number of SNPs was relatively small, the prediction accuracy of GS was also comparatively low.

Genomic selection using different cross-validation strategies

Different cross-validation strategies were adopted by partitioning the population into testing and validation sets with varying proportions to evaluate their impact on GS. Specifically, three schemes (3-fold, 5-fold, and 10-fold cross-validation) were implemented, corresponding to training set sizes of approximately 67%, 80%, and 90%, respectively. For the trait YZ_2020_FW, the results indicated that increasing the training set proportion from 67% to 80% and further to 90% led to a modest improvement in prediction accuracy; however, this increase was not statistically significant and was substantially lower than the improvement achieved through the use of top-associated SNPs selected via GWAS (Table 5, Supplementary Tables S3–S5).

Genomic selection using the weight-gBLUP model

The influence and association degree of different SNPs on traits are different; the association degree ($-\log_{10}(p\text{-value})$) can directly be taken as the weight to construct the weight-gBLUP model. It can be seen that after weighing, the GS model constructed using all SNPs has achieved a prediction accuracy of more than 0.8 for yield traits in various environments, which is higher than that of most SNP subsets (Fig. 5). Moreover, the variance of prediction accuracy of these models is very small, which is much better than the various models mentioned earlier.

Discussion

Phenotype and SNP number affect genomic selection

Phenotypic prediction determines the accuracy of subsequent GWAS and GS. However, the phenotypes of field plants are often influenced by the environment. To eliminate this influence, the BLUP was used to estimate the breeding values of plants and reflect the true genetic effects. In this study, the fresh weight yield of a total of five cuts from 2020 to 2022 was measured, and the BLUP values were calculated under various environments after considering the years and cuts for subsequent analysis. Using the BLUP value for GS, the prediction accuracy for most traits ranged from 0.10 to 0.28, which is notably low and consistent with previous studies^[29–31]. This might be due to the complexity of traits and the genetic characteristics of alfalfa tetraploid homology. However, in the research of various traits, there are also many traits with high prediction accuracy.

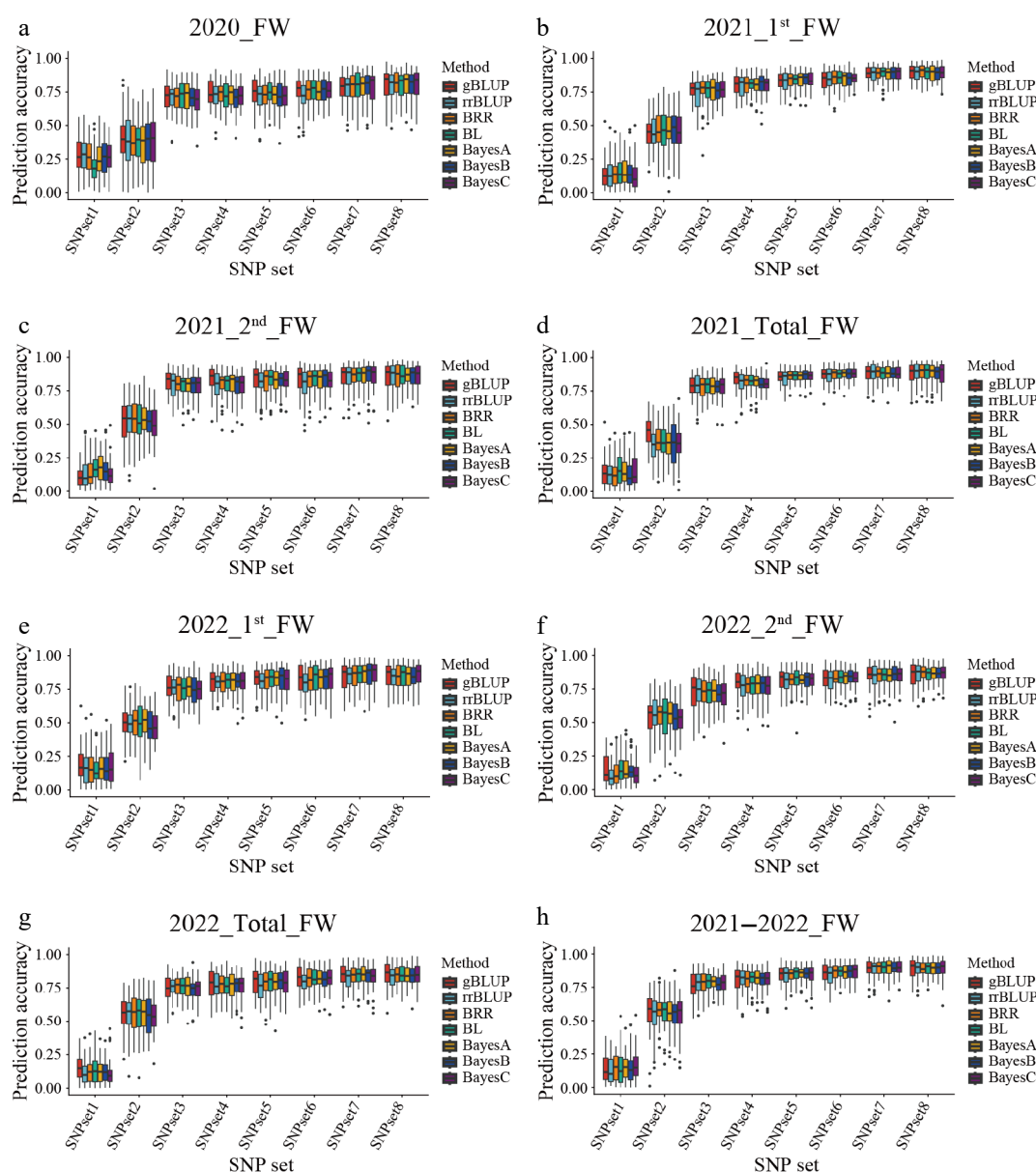


Fig. 3 Prediction accuracy of GS models for the FW traits using different SNP sets. (a)–(h) Represent the genomic selection (GS) results for different traits, corresponding to 2020_FW, 2021_1st_FW, 2021_2nd_FW, 2021_total_FW, 2022_1st_FW, 2022_2nd_FW, 2022_total_FW, and 2021–2022_FW, respectively. The different colors represent distinct GS models: gBLUP (red), rrBLUP (blue), BRR (green), BL (yellow), Bayes A (cyan), Bayes B (magenta), and Bayes C (purple).

Therefore, the phenotyping and analysis of alfalfa must be more rigorous and precise.

The heterozygosity of alfalfa is a major challenge in its genomic research. Therefore, when the reference genome was selected, although four copy sequences of each chromosome were obtained from cultivated alfalfa (cultivar XinJiangDaYe)^[21], only the longest copy was selected for application. A total of 211,610 high-quality SNPs were detected, while the influence of the number of SNPs on the prediction accuracy of alfalfa GS remained undetermined. In Li's research on alfalfa, it was found that the prediction accuracy of GS increased with the increase in the number of markers until the amount of missing data for each marker exceeded the 70%–80% limit^[32]. However, He et al. found that when the number of SNPs gradually increased from 1,000 to 161,170, the prediction accuracy fluctuated, but did not improve significantly^[29].

It was also found in this study that randomly selecting the number of SNPs did not significantly improve the prediction

accuracy. Conversely, when a small number of SNPs were randomly selected, the prediction accuracy was relatively low; as the number of randomly selected SNPs increased, the prediction accuracy improved accordingly, which aligns with Sipowicz et al.'s findings^[33]. Moreover, the prediction accuracy of GS in alfalfa using random SNPs was also relatively low (< 0.3); this might be because the selection of SNP subsets is random and does not take into account the effect values of SNPs on the selected traits. However, unlike previous reports, the plateau in this study occurred when the number of SNPs reached 5,000, which is considerably higher than the 500 or 1,000 reported previously. A similar phenomenon was observed in blueberries, where 5,000 markers yielded prediction ability comparable to that achieved with 86,000 markers^[34]. However, the influence of SNP numbers on GS is not entirely like this; Wang et al. found that the prediction accuracy using SNP subsets was generally slightly higher than that using all SNPs^[30].

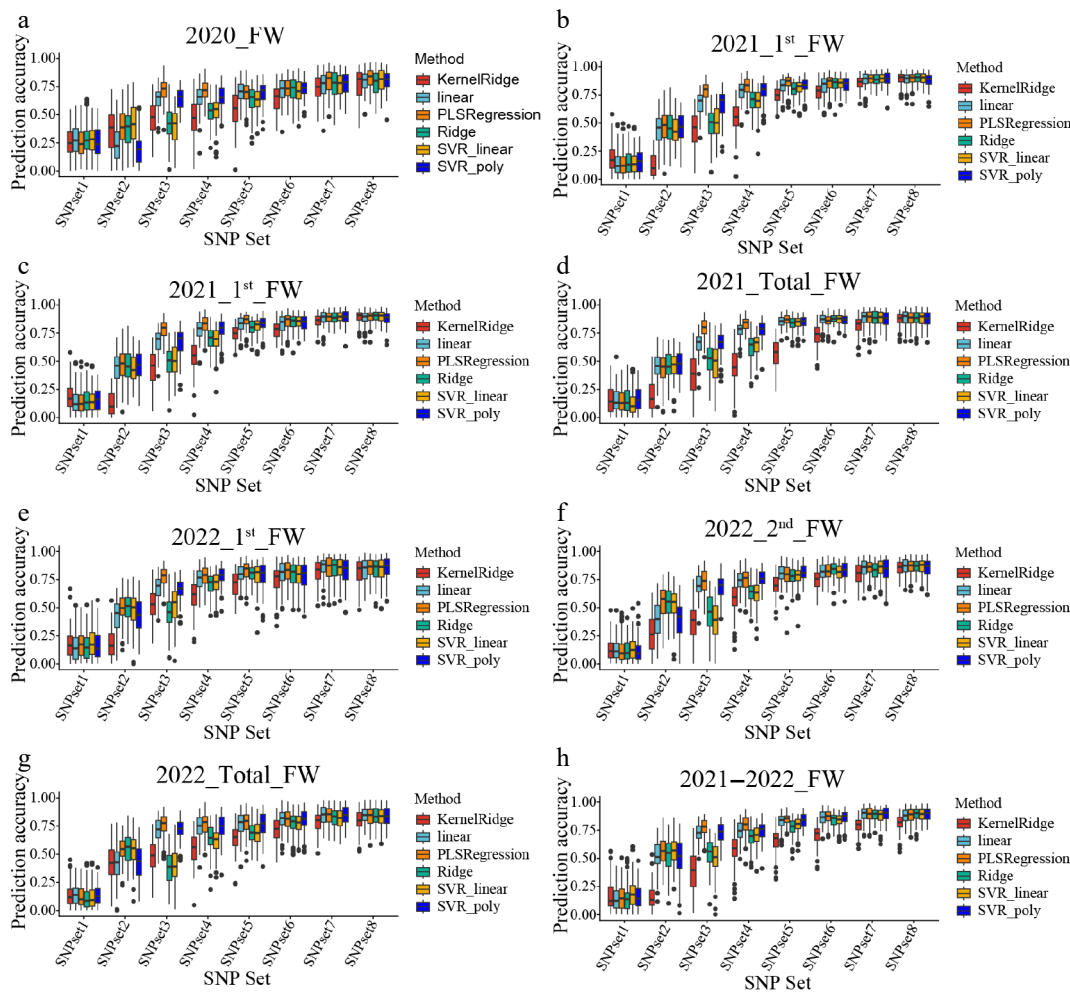


Fig. 4 The prediction accuracy of different machine learning models for the FW traits by using different SNP sets. (a)–(h) Represent the genomic selection (GS) results for different traits, corresponding to 2020_FW, 2021_1st_FW, 2021_2nd_FW, 2021_total_FW, 2022_1st_FW, 2022_2nd_FW, 2022_total_FW, and 2021–2022_FW, respectively. Different colors indicate different machine learning models: KernelRidge (red), linear (blue), PLSRegression (cyan), Ridge (yellow), SVR_linear (green), and SVR_poly (purple).

Table 5. The mean prediction accuracy of seven traditional models by using different SNP sets and validation strategies for the 2020_FW trait.

Strategy	SNP_Set1	SNP_Set2	SNP_Set3	SNP_Set4	SNP_Set5	SNP_Set6	SNP_Set7	SNP_Set8
3-fold	0.25	0.36	0.7	0.72	0.73	0.74	0.78	0.8
5-fold	0.26	0.37	0.71	0.73	0.73	0.75	0.79	0.8
10-fold	0.30	0.40	0.72	0.73	0.74	0.76	0.79	0.81

3-fold, 5-fold, and 10-fold denote k-fold cross-validation strategies.

Genomic selection combined with GWAS

The GWAS analysis identified 19 significant SNPs for FW traits, with several mapping to biologically relevant candidate genes. Notably, significant SNPs were located within or near genes encoding an LRR-repeat F-box protein (*MS.gene046331*) and an ankyrin repeat-containing protein (*MS.gene67840*), both implicated in the ubiquitin-proteasome system that regulates protein turnover in plant growth^[35,36]. Another significant SNP was identified in an exon of tryptophan synthase (*MS.gene027196*), a key enzyme in auxin biosynthesis^[37], while a promoter SNP was found in a casein kinase gene (*MS.gene072249*) involved in cell cycle regulation^[38]. These candidate genes, functionally linked to protein degradation, phyto-hormone synthesis, and cell cycle control, provide biological credibility to our GWAS results and represent potential targets for molecular breeding.

Building upon these GWAS findings, the contribution of different SNPs to traits varies, can be known, so when selecting SNPs, the degree of their influence on traits can be considered. Wang constructed a GS model using a subset of SNPs with the high mean-squared-estimated-marker effect and found that it could slightly improve the prediction accuracy of the alfalfa GS model^[30]. However, more people utilize the association information between SNPs and traits brought by GWAS^[17,39,40]. Through this association information, the top SNPs are selected as SNP subsets for the construction of GS models. The results show that for all traits, the prediction accuracy using SNP subsets is generally higher than that using all SNP^[30]. Compared with random SNP subsets, SNP subsets from GWAS studies significantly improve the prediction accuracy of the model^[41]. This is consistent with the results in this study. After using SNP subsets composed of different numbers of top SNPs, the prediction accuracy of various models approached or even reached

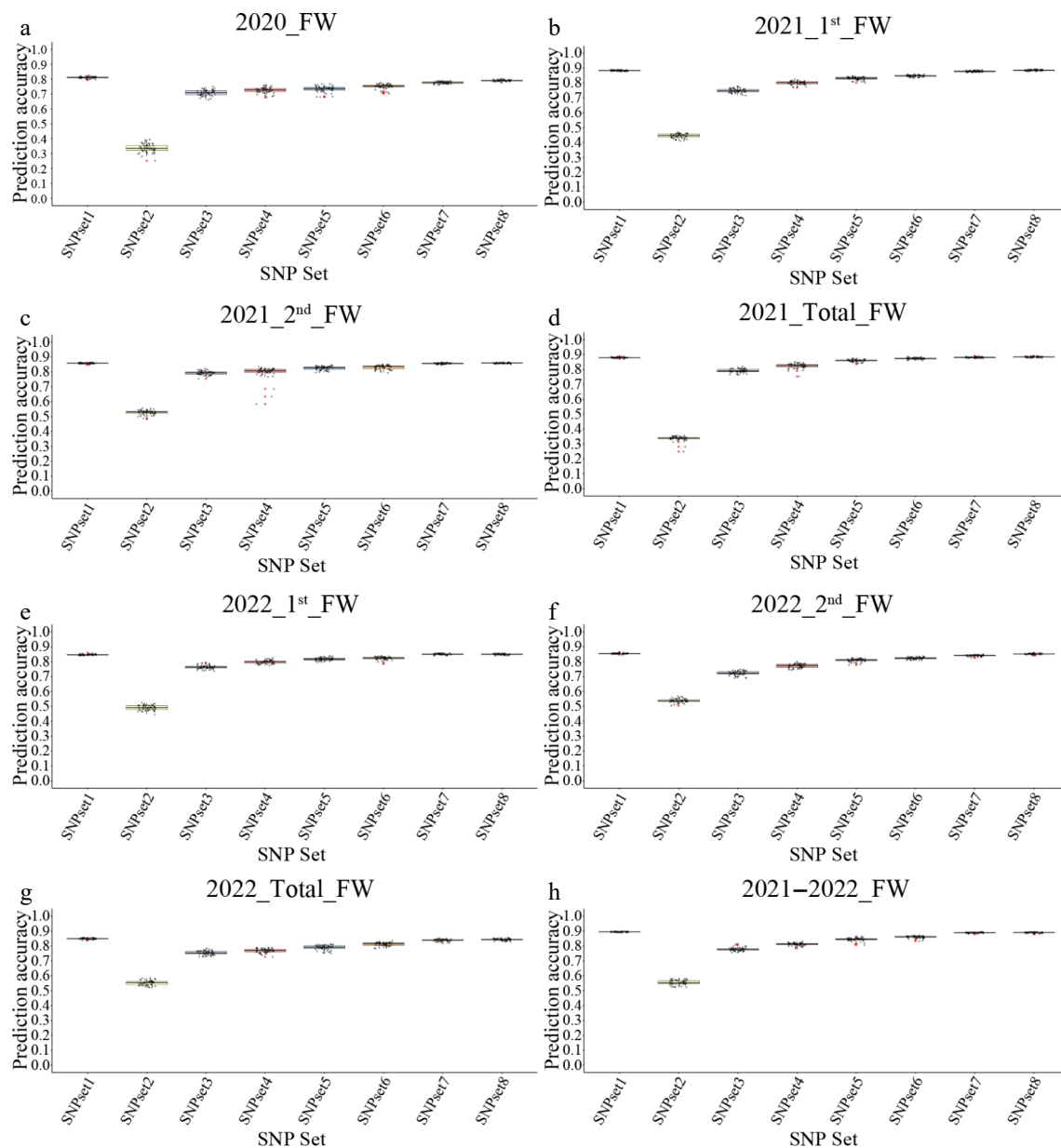


Fig. 5 The prediction accuracy of the weight-gBLUP models by using different SNP sets. (a)–(h) Represent the genomic selection (GS) results for different traits, corresponding to 2020_FW, 2021_1st_FW, 2021_2nd_FW, 2021_total_FW, 2022_1st_FW, 2022_2nd_FW, 2022_total_FW, and 2021–2022_FW, respectively.

0.9. This high prediction accuracy was not only observed in alfalfa^[18,41,42], but also found in various plants, such as maize (*Zea mays*)^[39], wheat (*Triticum aestivum* L.)^[41], and soybean (*Glycine max*)^[43]. Furthermore, when the proportion of the test population increased from a medium level (67%) to a higher level (80%, 90%), it was found that the prediction accuracy of each GS model in all SNP subsets tended to increase, which was consistent with the research results of other researchers^[17,39]. However, this increase is not significant compared with the increase in accuracy brought about by selecting top SNPs through GWAS.

Genomic selection using the weight-gBLUP model

The results of the research show that when using all SNPs, whether it is the traditional gBLUP, rrBLUP, and Bayes models, or the six models of machine learning, the difference in their prediction accuracy is not significant, and they are all relatively low. After selecting the top SNPs, the performance of various models on the SNP subset was similar, and the prediction accuracy began to

improve significantly, except that the improved speed of the machine learning KernelRidge model was not fast. Therefore, the improvement in the prediction accuracy of these models all utilizes the information from GWAS. Then, directly using the GWAS information of all SNPs to construct the weight-gBLUP model is also a good method to improve the prediction accuracy. After replacing the G matrix in the gBLUP model with the weighted G matrix constructed by $-\log_{10}(p\text{-value})$, the prediction accuracy of this model is also close to or even reaches 0.9. The number of SNPs with a p -value less than 0.05 is around 10,000, similar to the number of SNPs in SNPset8. However, the prediction accuracy result of constructing the weight-gBLUP model using SNPset8 is similar to that using all SNPs, both being close to 0.9. Combined with the results of the gBLUP model, it was found that these were basically consistent with the results of Medina et al.'s study on genomic prediction of alfalfa biomass yield^[13]. These results suggest that the weight-gBLUP model may be more suitable for the GS model of complex traits in alfalfa.

Conclusions

This study provides an effective strategy for breaking through the bottleneck of alfalfa yield breeding, which is to strategically integrate GS with GWAS. The main finding is that by using the information derived from GWAS, whether by selecting SNPs related to traits or by using association strength as genomic weights, the prediction accuracy of GS models for fresh weight yield in a half-sib population has been significantly improved. This work conducted a comparative evaluation of multiple modeling methods, and the results strongly demonstrated that the weighted gBLUP model, which uses the $-\log_{10}(p\text{-value})$ of GWAS as the weight for all SNPs, also achieved high prediction accuracy when all SNPs were utilized. This method effectively determines the priority of the information genome region without the need for arbitrary SNP subpopulation selection, providing a powerful yet streamlined alternative for breeding programs.

The practical significance of these findings is substantial. By achieving a prediction accuracy of 0.85–0.90, this method can more reliably select high-quality genotypes at the seedling stage, which has the potential to shorten the breeding cycle and accelerate the genetic gain of complex polygenic traits such as yield. Therefore, this study proposes the wgBLUP model as the optimal and effective strategy for implementing genomic selection in alfalfa semi-sibling family populations.

For future research, it will be valuable to validate this integrated GWAS-GS framework in larger and more diverse alfalfa populations and other important agronomic traits. In addition, the candidate genes identified by the GWAS provide new targets for molecular breeding, and in-depth research can help understand the genetic structure of yield.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Han Y, Yan Q, Zhang J; field experiment and phenotyping: Han Y, Ao B, Jing S, Chu B, Yan Q; data analysis and interpretation of results: Han Y, Zhang F, Xu P; preparation of manuscript: Han Y, Zhang J. All authors reviewed the results and approved the final version of the manuscript.

Data availability

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

Acknowledgments

This work was supported by the Biological Breeding Project (Grant No. 2022ZD04011); the National Key R&D Program of China (Grant No. 2022YFD1301002); the Major Science and Technology Project of Gansu Province (Grant No. 24ZD13NA014); the Gansu Province Modern Agricultural Industry Technology System (Grant No. GSARS-01); the 2025 Modern Cold and Arid Agriculture Seed Industry Research Project of the Department of Agriculture and Rural Affairs (Grant No. ZYGG-2025-19).

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper at (<https://www.maxapress.com/article/doi/10.48130/grares-0025-0028>)

Dates

Received 4 September 2025; Revised 5 October 2025; Accepted 14 October 2025; Published online 10 December 2025

References

- Bhattarai S, Wang H, Poudel HP, Biligetu B. 2024. Evaluating effectiveness of clonal plant selection of alfalfa (*Medicago sativa* L.) and sainfoin (*Onobrychis viciifolia* Scop.) in mixtures: mean performance and stability in a multi-environment trial. *Plant Breeding* 143:713–24
- Zhu H, Liu H, Liu Q, Wu Y, Xu Y, et al. 2025. Identification and response to abiotic stress of the alfalfa serine carboxypeptidase-like protein gene family. *Grass Research* 5:e022
- Shen C, Du H, Chen Z, Lu H, Zhu F, et al. 2020. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Molecular Plant* 13:1250–61
- Yang J, Yi J, Ma S, Wang Y, Song J, et al. 2024. Integrated physiological, metabolomic, and transcriptomic analyses elucidate the regulation mechanisms of lignin synthesis under osmotic stress in alfalfa leaf (*Medicago sativa* L.). *BMC Genomics* 25:174
- Wan W, Liu Q, Li K, Zhao K, Qi F, et al. 2025. Nitrogen fertilizer application for improving the biomass, quality, and nitrogen fixation of alfalfa (*Medicago sativa* L.) at different growth stages in a saline-alkali soil. *PeerJ* 13:e18796
- Wang W, Tian T, Li MY, Wang BZ, Mei FJ, et al. 2024. Carbon and nitrogen stoichiometry across plant–soil system accounts for the degradation of multi-year alfalfa grassland. *Frontiers in Plant Science* 15:1400261
- Annicchiarico P, Barrett B, Brummer EC, Julier B, Marshall AH. 2015. Achievements and challenges in improving temperate perennial forage legumes. *Critical Reviews in Plant Sciences* 34:327–80
- Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–29
- He F, Zhang F, Jiang X, Long R, Wang Z, et al. 2022. A genome-wide association study coupled with a transcriptomic analysis reveals the genetic loci and candidate genes governing the flowering time in alfalfa (*Medicago sativa* L.). *Frontiers in Plant Science* 13:913947
- Fu W, da Silva Linge C, Lawton JM, Gasic K. 2022. Feasibility of genomic prediction for brown rot (*Monilinia* spp.) resistance in peach. *Fruit Research* 2:2
- Xie C, Xu S. 1998. Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. *Heredity* 80:489–98
- Zhang D, Yang F, Li J, Liu Z, Han Y, et al. 2025. Progress and perspectives on genomic selection models for crop breeding. *Technology in Agronomy* 5:e006
- Medina CA, Hansen J, Crawford J, Viands D, Sapkota M, et al. 2025. Genome-wide association and genomic prediction of alfalfa (*Medicago sativa* L.) biomass yield under drought stress. *International Journal of Molecular Sciences* 26:608
- Medina CA, Hawkins C, Liu XP, Peel M, Yu LX. 2020. Genome-wide association and prediction of traits related to salt tolerance in autotetraploid alfalfa (*Medicago sativa* L.). *International Journal of Molecular Sciences* 21:3361
- Nazzicari N, Franguelli N, Ferrari B, Pecetti L, Annicchiarico P. 2024. The effect of genome parametrization and SNP marker subsetting on genomic selection in autotetraploid alfalfa. *Genes* 15:449
- Xu M, Jiang X, He F, Sod B, Yang T, et al. 2023. Genome-wide association study (GWAS) identifies key candidate genes associated with leaf size in alfalfa (*Medicago sativa* L.). *Agriculture* 13:2237
- Zhang F, Kang J, Long R, Li M, Sun Y, et al. 2022. Application of machine learning to explore the genomic prediction accuracy of fall dormancy in autotetraploid alfalfa. *Horticulture Research* 10:uhac225
- Medina CA, Kaur H, Ray I, Yu LX. 2021. Strategies to increase prediction accuracy in genomic selection of complex traits in alfalfa (*Medicago sativa* L.). *Cells* 10:3372
- Annicchiarico P, Nazzicari N, Brummer EC. 2016. Alfalfa genomic selection: challenges, strategies, transnational cooperation. In *Breeding in a*

- World of Scarcity*, eds. Roldán-Ruiz I, Baert J, Reheul D. Cham, Switzerland: Springer. pp.145–49 doi: [10.1007/978-3-319-28932-8_22](https://doi.org/10.1007/978-3-319-28932-8_22)
20. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297–303
 21. Chen H, Zeng Y, Yang Y, Huang L, Tang B, et al. 2020. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature Communications* 11:2494
 22. VanRaden PM. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91:4414–23
 23. Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* 4:250–55
 24. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. 2011. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186
 25. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182:375–85
 26. Park T, Casella G. 2008. The Bayesian Lasso. *Journal of the American Statistical Association* 103:681–86
 27. Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198:483–95
 28. Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, et al. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One* 5:e12648
 29. He X, Zhang F, He F, Shen Y, Yu LX, et al. 2022. Accuracy of genomic selection for alfalfa biomass yield in two full-sib populations. *Frontiers in Plant Science* 13:1037272
 30. Wang H, Bai Y, Biligetu B. 2024. Effects of SNP marker density and training population size on prediction accuracy in alfalfa (*Medicago sativa* L.) genomic selection. *The Plant Genome* 17:e20431
 31. Jiang X, Zeng X, Xu M, Li M, Zhang F, et al. 2025. The whole-genome dissection of root system architecture provides new insights for the genetic improvement of alfalfa (*Medicago sativa* L.). *Horticulture Research* 12:uhae271
 32. Li XH, Wei YL, Acharya A, Hansen JL, Crawford JL, et al. 2015. Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *Plant Genome* 8:eplantgenome2014.12.0090
 33. Sipowicz P, Andrade MHML, Filho CCF, Benevenuto J, Muñoz P, et al. 2025. Optimization of high-throughput marker systems for genomic prediction in alfalfa family bulks. *The Plant Genome* 18:e20526
 34. de Bem Oliveira I, Amadeu RR, Ferrão LFV, Muñoz PR. 2020. Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity* 125:437–48
 35. Hong MJ, Kim JB, Seo YW, Kim DY. 2021. Regulation of glycosylphosphatidylinositol-anchored protein (GPI-AP) expression by F-box/LRR-repeat (FBXL) protein in wheat (*Triticum aestivum* L.). *Plants* 10:1606
 36. Kane EI, Spratt DE. 2021. Structural insights into ankyrin repeat-containing proteins and their influence in ubiquitylation. *International Journal of Molecular Sciences* 22:609
 37. Mano Y, Nemoto K. 2012. The pathway of auxin biosynthesis in plants. *Journal of Experimental Botany* 63:2853–72
 38. Qu L, Wei Z, Chen HH, Liu T, Liao K, et al. 2021. Plant casein kinases phosphorylate and destabilize a cyclin-dependent kinase inhibitor to promote cell division. *Plant Physiology* 187:917–30
 39. Yan H, Guo H, Xu W, Dai C, Kimani W, et al. 2023. GWAS-assisted genomic prediction of cadmium accumulation in maize kernel with machine learning and linear statistical methods. *Journal of Hazardous Materials* 441:129929
 40. Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, et al. 2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116:395–408
 41. Alemu A, Brazauskas G, Gaikpa DS, Henriksson T, Islamov B, et al. 2021. Genome-wide association analysis and genomic prediction for adult-plant resistance to *Septoria tritici* blotch and powdery mildew in winter wheat. *Frontiers in Genetics* 12:661742
 42. Yu LX, Liu X, Boge W, Liu XP. 2016. Genome-wide association study identifies loci for salt tolerance during germination in autotetraploid alfalfa (*Medicago sativa* L.) using genotyping-by-sequencing. *Frontiers in Plant Science* 7:956
 43. Singer WM, Shea Z, Yu D, Huang H, Rouf Mian MA, et al. 2022. Genome-wide association study and genomic selection for proteinogenic methionine in soybean seeds. *Frontiers in Plant Science* 13:859109



Copyright: © 2025 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.