agronomy

MDPI

Article

Genomic Selection in Alfalfa Across Multiple Ploidy Levels: A Comparative Study Using Machine Learning and Bayesian Methods

Xiaoyue Zhu, Ruixin Zhang, Tianxiang Zhang, Changhong Guo and Yongjun Shu *

Key Laboratory of Molecular Cytogenetics and Genetic Breeding of Heilongjiang Province, College of Life Science and Technology, Harbin Normal University, Harbin 150025, China; zhuxiaoyue2001@126.com (X.Z.); zrxin2001@126.com (R.Z.); hsdztx@stu.hrbnu.edu.cn (T.Z.); kaku3008@hrbnu.edu.cn (C.G.)
* Correspondence: syjun2003@126.com or shuyj@hrbnu.edu.cn; Tel.: +86-451-88060576

Abstract: Agronomic traits and quality traits of alfalfa are of great importance to the feed industry. Genomic selection (GS) based on genotyping-by-sequencing (GBS) data, if it achieves moderate to high accuracy, has the potential to significantly shorten breeding cycles for complex traits and accelerate genetic progress. This study aims to investigate the effect of different reference genomes on the prediction accuracy of genomic selection. A total of 11 Bayesian and machine learning models and nine different reference genomes were used to conduct genomic selection on five traits in 385 alfalfa accessions. The accuracy of GS was evaluated using five-fold cross-validation, based on the correlation between genomic estimated breeding values (GEBVs) and estimated breeding values (EBVs). For the five traits, it was found that traits with high heritability exhibited significantly higher prediction accuracy. The prediction accuracy fluctuated minimally across different reference genomes, with the diploid genome showing relatively higher accuracy. For two high-heritability traits, fall dormancy and plant height, predictions were made after SNP density reduction, and it was observed that density had little effect on prediction accuracy. However, for the fall dormancy trait in the diploid genome, more than half of the models showed regular fluctuations, with prediction accuracy increasing as SNP density increased. In conclusion, this study provides a theoretical basis for precision breeding of alfalfa and other polyploid crops by combining different reference genomes and models, and offers important guidance for optimizing future genomic selection strategies.

Keywords: genomic selection; alfalfa; Bayesian model; machine learning; SNP density; heritability

Citation: Zhu, X.; Zhang, R.; Zhang, T.; Guo, C.; Shu, Y. Genomic Selection in Alfalfa Across Multiple Ploidy Levels: A Comparative Study Using Machine Learning and Bayesian Methods. *Agronomy* **2024**, *14*, 2768. https://doi.org/10.3390/agronomy14122768

Academic Editors: Ruijun Qin, Johnny Li and Tajamul Hussain

Received: 25 October 2024
Revised: 17 November 2024
Accepted: 20 November 2024
Published: 21 November 2024

1. Introduction

Alfalfa (*Medicago sativa* L.), a widely cultivated perennial leguminous forage crop, plays a critical role in global agricultural production, offering substantial economic and ecological benefits [1]. It is not only a primary source of high-quality feed but is also widely used in agricultural ecosystems due to its nitrogen-fixing ability and soil-improvement effects [2]. In recent years, with the increasing global demand for efficient and sustainable agriculture, the breeding and improvement of alfalfa have become increasingly important. Nevertheless, the complex genetic background of alfalfa, along with its polyploid nature (both tetraploid and diploid), and the diverse environmental conditions it encounters present significant challenges to traditional breeding approaches targeting its key traits [3,4]. Therefore, there is an urgent need to introduce new breeding strategies into alfalfa breeding programs to accelerate genetic gains in target traits, thereby meeting the growing demand for feed production.

Genomic selection (GS) offers new possibilities for alfalfa breeding by combining phenotypic data and genome-wide marker information to predict the breeding value of individuals, thereby reducing breeding cycles and improving selection efficiency [5]. GS methods based on genome-wide markers have made some progress in plant breeding [2,6].

For example, genomic selection using Bayesian methods and different cross-validation techniques has been applied in crops like soybeans, rice, and maize, with prediction accuracies reaching up to 0.9 for certain traits [7]. Additionally, machine learning methods have recently been used to predict crop traits. For alfalfa yield prediction, machine learning models can leverage weather data, historical yield information, and planting dates to improve accuracy and generate more reliable forecasts [8–10]. The study by Zhang et al. [11] showed that combining machine learning with GWAS-associated markers significantly improved the prediction accuracy of genomic selection for complex traits such as fall dormancy in polyploid crops, achieving 64.1% accuracy. Moreover, the choice of SNP density also has a significant impact on the accuracy of genomic selection. Kriaridou et al. [12] demonstrated that by using low-density SNP panels combined with genotype imputation techniques, it is possible to maintain a prediction accuracy similar to that of high-density panels while reducing costs. This approach provides a more cost-effective solution for large-scale implementation of genomic selection, further promoting its application in agricultural breeding programs, including alfalfa.

However, the application of genomic selection in alfalfa still faces numerous challenges. Alfalfa is a typical tetraploid plant ($2n = 4x = 32$), which makes one of the biggest challenges in genomic selection how to handle its complex genomic structure. Tetraploid plants have four genome copies, and genomic selection often requires one to accurately integrate the genotypic and phenotypic information of these copies [13]. There can be significant variation between these copies, which complicates the association between genetic markers and phenotypes. Additionally, the alfalfa genome is large and highly diverse, which presents another issue in genomic selection: how to efficiently manage and analyze large-scale genomic data. The alfalfa genome is large and contains a substantial number of repetitive sequences and transposons, which complicates whole-genome sequencing and SNP marker identification [14]. Therefore, genomic selection research for alfalfa and other polyploid crops remains an area that urgently requires further exploration.

Given the polyploid genome and complex genetic architecture of alfalfa, selecting an appropriate reference genome is essential for ensuring the accuracy of genomic selection outcomes [15]. The effectiveness of genome-wide markers, and consequently the accuracy of trait prediction, can be influenced by the choice of reference genome. The diversity of the reference genome, its phylogenetic relationship to the target population, and the quality of its assembly are key factors that can significantly influence the outcomes of genomic selection. For instance, Jia et al. [16] utilized the genome of *Medicago sativa* to predict 25 agronomic and quality traits, applying three Bayesian statistical methods—BayesA, BayesB, and BayesCπ—and demonstrated varying levels of prediction accuracy across traits. Similarly, Medina et al. [17] investigated salt stress tolerance in alfalfa, using *Medicago truncatula* as the reference genome. Through genome-wide association studies (GWASs) and genomic selection (GS) methods, they identified SNPs associated with salt stress and employed machine learning models such as support vector machines and random forests to enhance prediction accuracy. Annicchiarico et al. [18] compared genomic selection using three different reference genomes for alfalfa: a simulated genome, a diploid genome (*M. truncatula*), and a tetraploid genome. Their results indicated that prediction accuracy followed the order mock reference genome > *M. truncatula* > *M. sativa*, providing valuable insights into genomic selection at the polyploid level. However, there have been no comprehensive studies to date that directly compare the effects of these reference genomes on prediction accuracy. Therefore, further research on the impact of different reference genomes on genomic prediction accuracy is of critical importance for breeding practices.

This study aims to evaluate the performance of various reference genomes in alfalfa genomic selection by exploring the predictive ability of different statistical models and SNP densities for high-heritability traits such as fall dormancy and plant height. The research investigates how reference genomes influence genomic selection outcomes, providing theoretical guidance for precision breeding of alfalfa and offering insights into genomic selection strategies for other polyploid crops.

2. Materials and Methods

2.1. Data Collection

The phenotypic data used in this study were derived from the research conducted by Pégard et al. (2023) [19]. This research carried out phenotypic evaluations on 400 alfalfa cultivars, assessing traits such as fall dormancy and flowering date, with measurements taken in two locations, France and Serbia. The experiment consisted of 440 plots arranged in 44 columns and 10 rows, using an augmented block design with four incomplete blocks. The micro-environmental spatial effects were subtracted from the observed phenotypes to obtain spatially adjusted phenotypes [20]. After filtering out duplicates, a total of 385 phenotypic data points were used in this study. The traits included flowering date (FD) and autumn dormancy, measured by various parameters in autumn (dormancy (D), forage dry matter yield (F-DMY), plant height (PH), and speed of elongation (SE)) over two years, 2019 (X19.X) and 2020 (X20.X), in two locations: Lusignan (.L) in France and Novi Sad (.N) in Serbia. The phenotypic data used in this study were measured in 2019 in Lusignan and include D19.L, F.DMY19.L, PH19.L, SE19.L, and FD.L.

The GBS sequences used in this study are available in the NCBI SRA under BioProject PRJNA961940, which includes 1012 individuals [19]. To explore the differences in genomic selection at different ploidy levels, various alfalfa reference genomes were collected, as shown in Table 1. The first column lists the genome labels used in this study, the second column provides their Latin names, and the third column cites the reference from which the genome data were sourced. The collected genomes include seven previously reported genomes, as well as two haploid genomes derived from Xinjiang Daye and Zhongmu No. 4 alfalfa in this study, resulting in a total of nine distinct genomes.

Table 1. This table summarizes the genome assembly data for various alfalfa species and cultivars, including both tetraploid and diploid forms.

Index	Latin Name	Ploidy	References
MtA17	*Medicago truncatula*	diploid	[14]
MsDip	*Medicago sativa* spp. *caerulea*	diploid	[21]
MsJHC	*Medicago polymorpha*	tetraploid	[22]
MruH	*Medicago ruthenica*	tetraploid	[23]
MsZM1	*Medicago sativa* Zhongmu-1	tetraploid	[24]
MsXJDY	*Medicago sativa* XinjiangDaye	tetraploid	[14]
MsZM4	*Medicago sativa* Zhongmu-4	tetraploid	[25]
MsXJDY-1H	XinjiangDaye_haploid	haploid	[14]
MsZM4-1H	Zhongmu-4_haploid	haploid	[25]

2.2. SNP Identification and Data Processing

In order to detect single-nucleotide polymorphisms (SNPs), sequencing reads were first aligned to the reference genome using the BWA-MEM algorithm within BWA software (version 0.7.17), producing BAM-format alignment files [26]. Subsequently, SAMtools was employed to sort and index these files, optimizing the efficiency of downstream analyses [27]. Additionally, samtools flagstat can be used to generate statistics for BAM files, including alignment quality, unique mapping rate, and duplication rate [28]. SNP identification was then conducted using BCFtools (version 1.10.2) [27]. The process began by using the bcftools mpileup function to aggregate alignment data, producing a raw VCF file that contained nucleotide variations and the corresponding read depth at each site [29]. Following this, SNPs were identified with the bcftools call function, applying a multi-allelic model to account for variation in regions with multiple alleles. After SNP calling, the dataset was subjected to a filtering process, retaining only high-confidence SNPs by excluding variants with a Phred quality score below 20 or a read depth lower than 10. Next, the VCF dataset was refined using PLINK software (version 1.90b7.2) by removing loci with a genotype missing rate greater than 20% or a minor allele frequency (MAF) less than 0.05, ensuring that only bi-allelic sites remained for subsequent analyses [30]. To

impute any missing genotypes, Beagle software (version 5.1) was utilized, resulting in a comprehensive SNP dataset for further use in analysis [31]. In this study, to investigate the impact of SNP density on prediction accuracy, we used the—thin parameter in vcftools (version 0.1.17) to perform thinning, extracting one SNP every 20 k, 50 k, and 100 k bases, resulting in three different SNP densities [32].

2.3. Genomic Prediction with Bayes Models

To perform genomic selection, several Bayesian models were employed to estimate the marker effects on phenotypes. These models included BayesA, BayesB, BayesC, Bayesian Lasso (BL), and Bayesian Ridge Regression (BRR), all implemented using BGLR software (version 4.4.1) [33]. The general linear model used for genomic prediction is

$$y = X\beta + Zg + \epsilon \quad (1)$$

where y is the vector of phenotypic values, X is the design matrix for fixed effects, β is the vector of fixed effect coefficients, Z is the genotype matrix (markers), g is the vector of marker effects, and ϵ is the residual error, which is assumed to follow a normal distribution, $\epsilon \sim N(0, \sigma_\epsilon^2)$. The posterior distribution $p(g|y)$ for marker effects g is derived using Bayes' theorem:

$$p(g|y) \propto p(y|g) \cdot p(g) \quad (2)$$

where $p(g)$ represents the prior distribution for marker effects, which differs among the models:

BayesA: assumes each marker effect g_i follows a normal distribution with a marker-specific variance of

$$g_i \sim N\left(0, \sigma_j^2\right) \quad (3)$$

The variance σ_j^2 for each marker is drawn from an inverse Chi-Square prior distribution. This allows for different shrinkage across markers. The prior distribution for marker variance is

σ_j^2~Inverse Chi-Square (ν,S^2)

BayesB introduces sparsity by assuming that most marker effects are zero, and only a small proportion π of markers have non-zero effects. The prior for marker effects is

$$g_j \sim (1-\pi)\delta_0 + \pi N\left(0, \sigma_j^2\right) \quad (4)$$

where π is the proportion of non-zero markers and δ_0 represents the Dirac delta function for zero effects. Non-zero marker effects g_i are drawn from a normal distribution with marker-specific variances.

BayesC is similar to BayesB, but assumes that all non-zero marker effects share a common variance, σ_g^2. The prior for marker effects is

$$g_j \sim (1-\pi)\delta_0 + \pi N\left(0, \sigma_g^2\right) \quad (5)$$

Here, σ_g^2 is shared across all non-zero markers, simplifying the variance structure.

Bayesian Lasso (BL) assumes that marker effects follow a Laplace (double-exponential) distribution, which imposes a Lasso-like shrinkage:

$$p(g_j|\lambda) = \frac{\lambda}{2}\exp(-\lambda|g_j|) \quad (6)$$

The parameter λ controls the degree of shrinkage, encouraging sparsity in the marker effects by shrinking most effects towards zero.

Bayesian Ridge Regression (BRR) assumes that all marker effects follow a normal distribution with a common variance:

$$g_j \sim N\left(0, \sigma_g^2\right) \tag{7}$$

This model applies uniform shrinkage to all marker effects, assuming that many loci contribute small effects to the trait.

For all Bayesian methods, the analyses were performed using the BGLR package in the R environment (version 4.4.1) [33]. The MCMC chain was run for 50,000 iterations, with the first 1000 iterations discarded as burn-in. Additionally, 5-fold cross-validation was conducted during the genomic selection process, with each model repeated 50 times.

2.4. Genomic Prediction with Machine Learning Models

2.4.1. Ridge Regression

Ridge Regression is a linear regression model with L2 regularization to prevent overfitting. The regularization term penalizes large coefficients, encouraging the model to generalize better. The parameter α controls the strength of regularization, where larger α values lead to stronger regularization.

$$\min_w \| y - Xw \|_2^2 + \alpha \| w \|_2^2 \tag{8}$$

where y is the target variable, X is the input matrix, w represents the model coefficients, and α is the regularization strength.

2.4.2. Kernel Ridge Regression

Kernel Ridge Regression combines Ridge Regression with kernel methods. It allows the model to capture non-linear relationships by mapping input data into a higher-dimensional space using a kernel function. The regularization prevents overfitting by penalizing large coefficients.

$$\min_\alpha \| y - K\alpha \|_2^2 + \lambda \| \alpha \|_2^2 \tag{9}$$

where K is the kernel matrix, α represents the weights, and α is the regularization parameter.

2.4.3. Partial Least Squares Regression (PLS Regression)

PLS regression is used when there are many variables and multicollinearity. It finds components that explain the most variance in both the input variables and the output, making it useful for high-dimensional data.

PLS regression does not have a simple formula but operates by finding a set of latent factors that maximize the covariance between input data X and response Y. It projects X and Y onto new spaces while preserving the relationship between them:

$$\begin{aligned} T &= XW \\ Y &= TQ + E \end{aligned} \tag{10}$$

where T is the latent variable, W and Q are loading matrices, and E represents residuals.

2.4.4. Support Vector Regression–Linear Kernel (SVR_linear)

Support Vector Regression (SVR) with a linear kernel is a regression method that aims to minimize prediction errors while maximizing the margin around the predicted values. The parameter C balances between fitting the training data and maintaining a large margin.

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum \xi_i \tag{11}$$

Subject to:

$$\begin{aligned} y_i - (w \cdot x_i + b) &\leq \epsilon + \xi_i \\ (w \cdot x_i + b) - y_i &\leq \epsilon + \xi_i \end{aligned} \tag{12}$$

where ξ are slack variables, C controls the penalty for errors, and ϵ defines the margin of tolerance.

2.4.5. Support Vector Regression-Polynomial Kernel (SVR_poly)

Support Vector Regression with a polynomial kernel extends linear SVR by using a polynomial kernel to capture non-linear relationships in the data. It maps input data into a higher-dimensional space where they can be linearly separated, making it suitable for complex data.

$$K(x_i, x_j) = (x_i \cdot x_j + r)^d \tag{13}$$

where r is a constant, d is the degree of the polynomial, and K is the kernel function.

2.4.6. Linear Regression

Linear regression is one of the simplest regression models. It assumes a linear relationship between the input variables and the output variable and finds the best-fit line by minimizing the sum of squared errors between predicted and actual values.

$$y = Xw + b \tag{14}$$

where y is the predicted value, X is the input matrix, w is the model coefficient, and b is the intercept.

The machine learning models used in this study were implemented by writing Python scripts and built using the Python package sklearn. Each model underwent 5-fold cross-validation and was run 50 times [11].

2.5. Heritability and Phenotypic Variance Explained

Heritability is a measure of the proportion of phenotypic variance that can be attributed to genetic variation. It reflects the contribution of genetic factors to the total variation observed in a particular trait. In this study, the contribution of all genetic effects to phenotypic variance, referred to as broad-sense heritability, was assessed as follows:

$$h^2 = \frac{\sigma_U^2}{\sigma_U^2 + \sigma_E^2} \tag{15}$$

Here, σ_U^2 represents the estimated genetic variance in the model (calculated from the effects of the genotype matrix, also known as additive genetic variance), and σ_E^2 represents the estimated residual variance in the model (the non-genetic component of the phenotypic variance).

To evaluate the predictive power of the model and assess the difficulty of predicting traits, the PVE (Phenotypic Variance Explained) value was also calculated in this experiment. The calculation method is as follows:

$$PVE = \frac{Var(X_{train} \cdot \hat{\beta})}{Var(y_{train})} \tag{16}$$

Here, X_{train} represents the genotype matrix in the training set, $\hat{\beta}$ denotes the regression coefficients (genetic effects) obtained from the model fitting, $Var(X_{train})$ is the variance in genetic effects in the training set, and $Var(y_{train})$ is the total variance in the phenotypic data in the training set.

All of the above methods were implemented using the BGLR package in the R environment [34]. To reduce randomness and avoid bias from chance occurrences, the calculations were repeated 100 times, and the average value was taken.

2.6. Cross-Validation and Genomic Prediction Accuracy

In this study, the Pearson correlation coefficient was used to assess the linear relationship between the true values and the predicted values from different models, which reflects the prediction accuracy [35]. A 50 × 5 CV (5-fold cross-validation repeated 50 times, totaling 250 trials) was conducted. The formula for calculating the correlation coefficient is as follows:

$$r = \frac{\sum\left(y_{true} - \overline{y}_{true}\right)\left(y_{pred} - \overline{y}_{pred}\right)}{\sqrt{\sum\left(y_{true} - \overline{y}_{true}\right)^2 \cdot \sum\left(y_{pred} - \overline{y}_{pred}\right)^2}} \quad (17)$$

Here, y_{true} represents the actual values (true phenotypic values), y_{pred} represents the predicted values (phenotypic values predicted by the model), and $\overline{y}_{true}$ and $\overline{y}_{pred}$ are the mean values of the actual and predicted values, respectively.

3. Results

3.1. The Impact of Different Reference Genomes on Mapping Rates

The properly paired rate refers to the proportion of paired-end reads that are correctly paired and mapped to the reference genome, indicating the success rate of read alignment and the quality of genome assembly. Figure 1 illustrates the distribution of properly paired rates across various reference genomes, including MtA17, MsDip, MsJHC, MruH, MsZM1, MsXJDY, MsZM4, MsXJDY-1H, and MsZM4-1H. The vertical axis represents the percentage of properly paired reads for each reference genome, providing a measure of the efficiency and accuracy of read pairing relative to the genome in use. A violin plot is employed to visualize the distribution of properly paired rates for each genome.

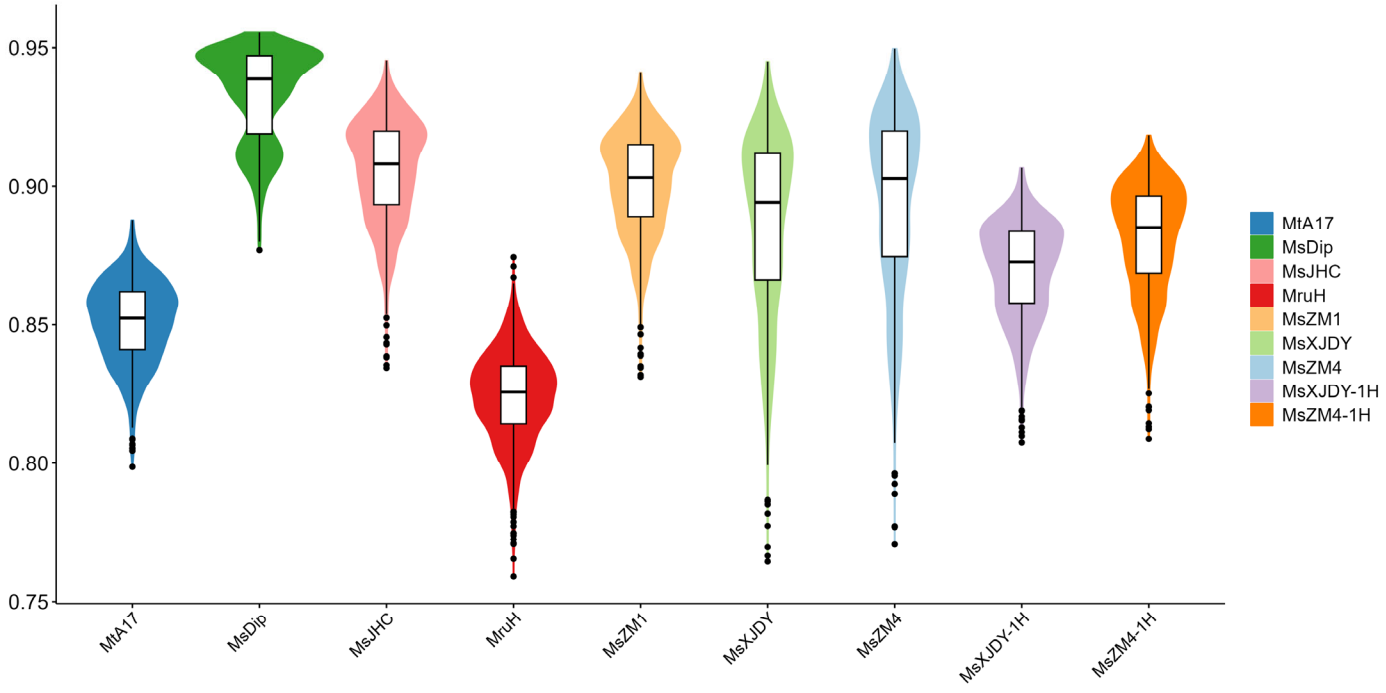

Figure 1. Violin plot of properly paired rates for different reference genomes, including MtA17, MsDip, MsJHC, MruH, MsZM1, MsXJDY, MsZM4, MsXJDY-1H, and MsZM4-1H. The plot illustrates the distribution of properly paired rates for each genome, with the width representing the density of data points at different paired rate intervals.

The figure demonstrates that the correct SNP pairing rates, from highest to lowest, follow the order of MsDip > MsJHC > MsZM4 > MsZM1 > MsZM4-1H > MsXJDY > MsXJDY-1H > MtA17 > MruH. Notably, MsDip exhibits the highest properly paired rate,

indicating superior genome assembly quality and a high degree of genetic similarity with the experimental samples. Therefore, MsDip is recommended as the primary reference genome for alfalfa research. In contrast, MtA17, despite being a model species for alfalfa, shows a relatively lower SNP pairing efficiency. This suggests that although the genome assembly quality of MtA17 is relatively good, there are significant genetic differences between this genome and the experimental samples, making it less suitable for the current study's needs. In the comparison between these two diploids, the correct pairing rate of MsDip is significantly higher than that of MtA17, possibly due to the large difference in their genome sizes, as well as the higher proportion of transposable elements in the MsDip genome. MruH performs the worst, with a low and broadly distributed pairing rate, potentially due to substantial genetic divergence or a more complex genome structure relative to the experimental samples. Additionally, among all the genomes, MsDip shows the fewest outliers, while MruH has the most outliers, with a wide distribution. The performance of MsXJDY and its haploid version is relatively poor, with notable fluctuations in the data, indicating that the pairing efficiency for these genomes is particularly low in certain samples. These variations among the reference genomes reflect differences in genome complexity, genetic diversity, and their relevance to the experimental samples. These differences between the reference genomes may reflect variations in diversity, genome complexity, and genetic relatedness to the experimental subjects. These findings provide a basis for selecting appropriate reference genomes for alfalfa genomic studies and genomic selection breeding.

3.2. Comparison of Heritability for Five Traits Across Different Reference Genomes

To assess the impact of different reference genomes on heritability estimates, the BGLR package was employed to calculate the heritability of five traits across various reference genomes, as depicted in Figure 2. The x-axis represents the different reference genomes, while the y-axis indicates heritability values. Each color corresponds to one of the five traits: D19.L, F.DMY19.L, PH19.L, SE19.L, and FD.L. As seen in the figure, the heritability for the D19.L trait ranges from 0.76 to 0.78, while for the F.DMY19.L trait, it spans from 0.64 to 0.67. For the PH19.L trait, heritability also falls within the 0.76 to 0.78 range, whereas the SE19.L trait ranges from 0.64 to 0.58, and the FD.L trait has a heritability range of 0.58 to 0.59. Among all genomes, the heritability of D19.L and F.DMY19.L traits is relatively high, generally exceeding 0.7, indicating a strong performance in heritability across various genes for these models. In contrast, FD.L shows comparatively lower heritability. The heritability performance across different traits in each genome remains relatively stable; D19.L and F.DMY19.L traits, in particular, maintain high heritability levels across all genomes, suggesting that these models possess consistent predictive ability across different genes. The results indicate that the choice of reference genome has minimal influence on heritability, as heritability across most genomes shows little fluctuation, except for certain traits in MsXJDY and MsJHC, which exhibit slightly lower values compared to other genomes.

3.3. Genomic Selection Prediction Accuracy

To evaluate the effects of different reference genomes and prediction models on prediction accuracy, 11 commonly used genomic selection models were employed, including BayesA, BayesB, BayesC, BL (Bayesian Lasso), BRR (Bayesian Ridge Regression), Kernel Ridge, linear regression, PLS regression, Ridge Regression, and Support Vector Regression (SVR), with both linear and polynomial kernels. These models were applied to genomic selection for five traits: D19.L, F.DMY19.L, PH19.L, SE19.L, and FD.L. Pearson correlation coefficients were used to assess the prediction accuracy, as depicted in Figures 3–7, as well as a distribution plot of the results repeated 50 times. The x-axis represents different reference genomes, while the y-axis displays the prediction accuracy achieved by each model in genomic selection. An asterisk marks the top three genomes in terms of average prediction accuracy for each model.

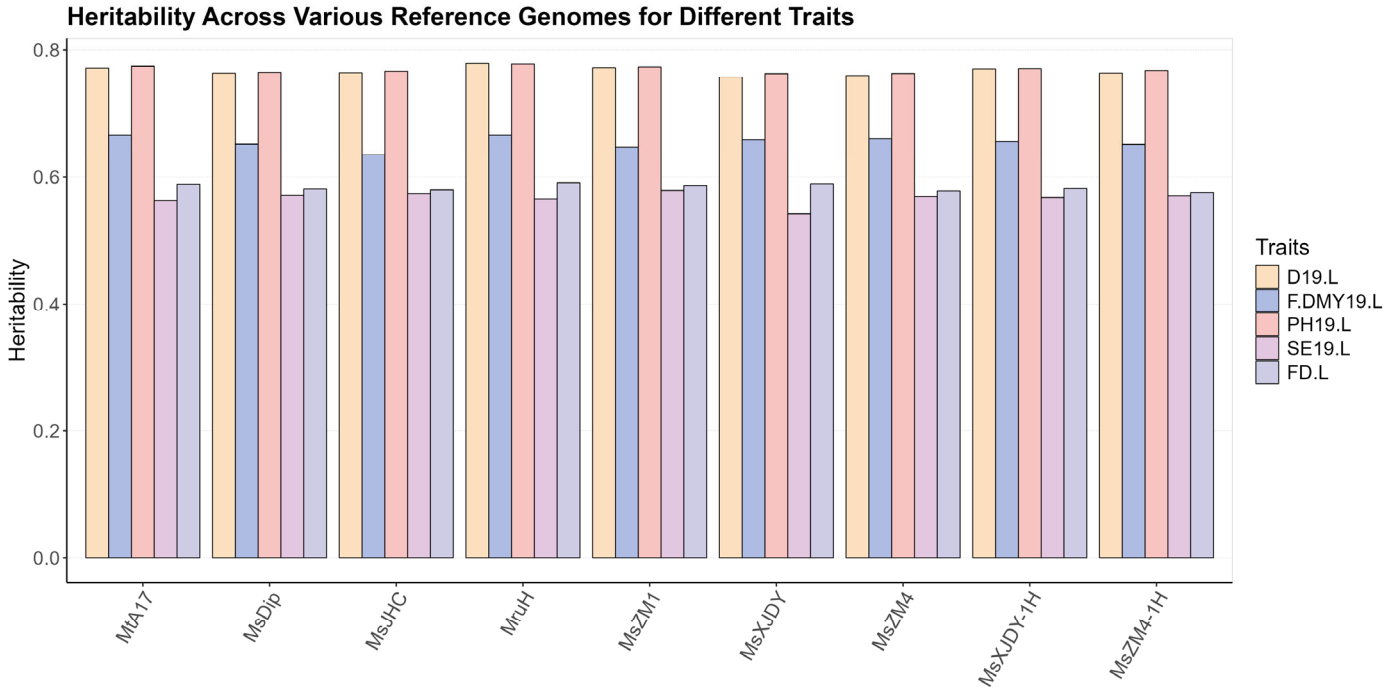

Figure 2. Heritability distribution of five traits across different reference genomes. The x-axis represents various reference genomes, including MtA17, MsDip, MsJHC, MruH, MsZM1, MsXJDY, MsZM4, MsXJDY-1H, and MsZM4-1H. Each group contains five traits (D19.L, F.DMY19.L, PH19.L, SE19.L, and FD.L), and the y-axis represents the heritability, ranging from 0.0 to 0.8. The bar chart shows the heritability differences for each combination of genome and trait, with colors indicating different trait types.

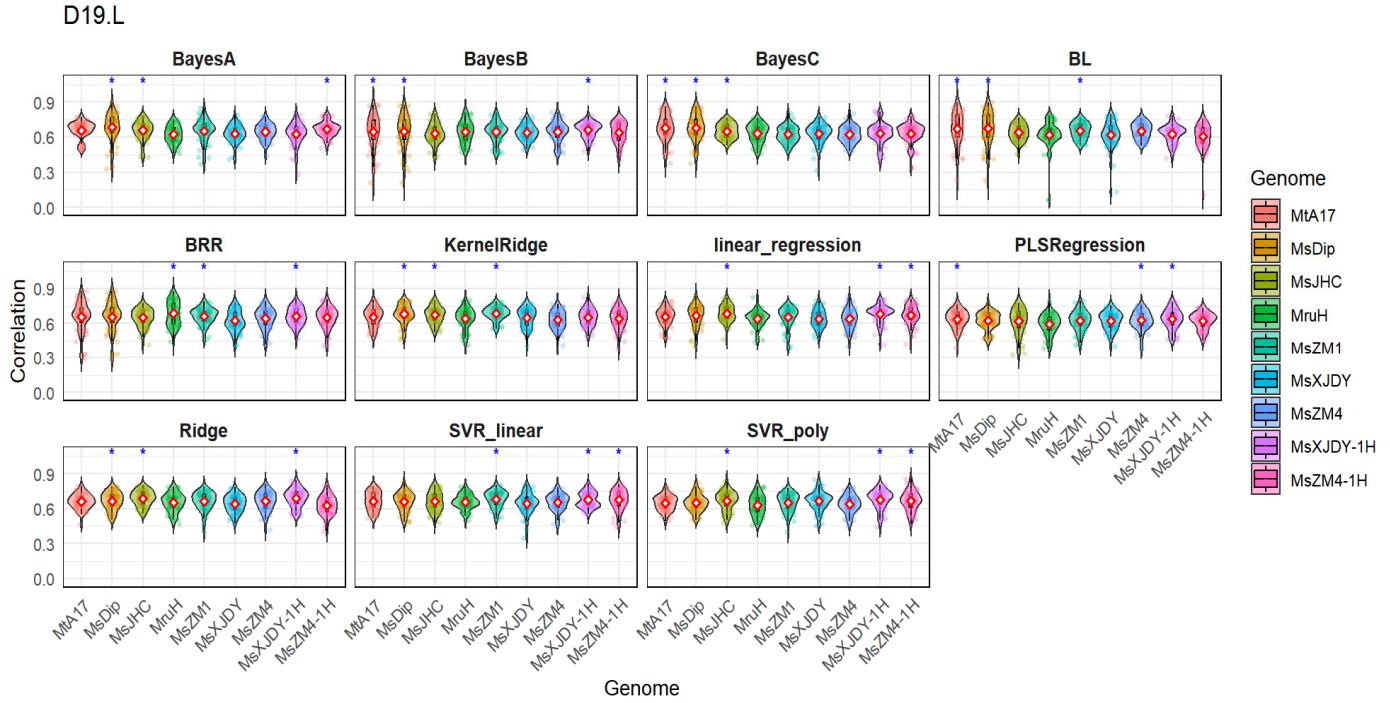

Figure 3. Comparison of prediction accuracy for the D19.L trait across different reference genomes. The plot displays the prediction accuracy distribution across 9 different reference genomes for 11 models. An asterisk (*) indicates the top three models based on the average prediction accuracy for each model.

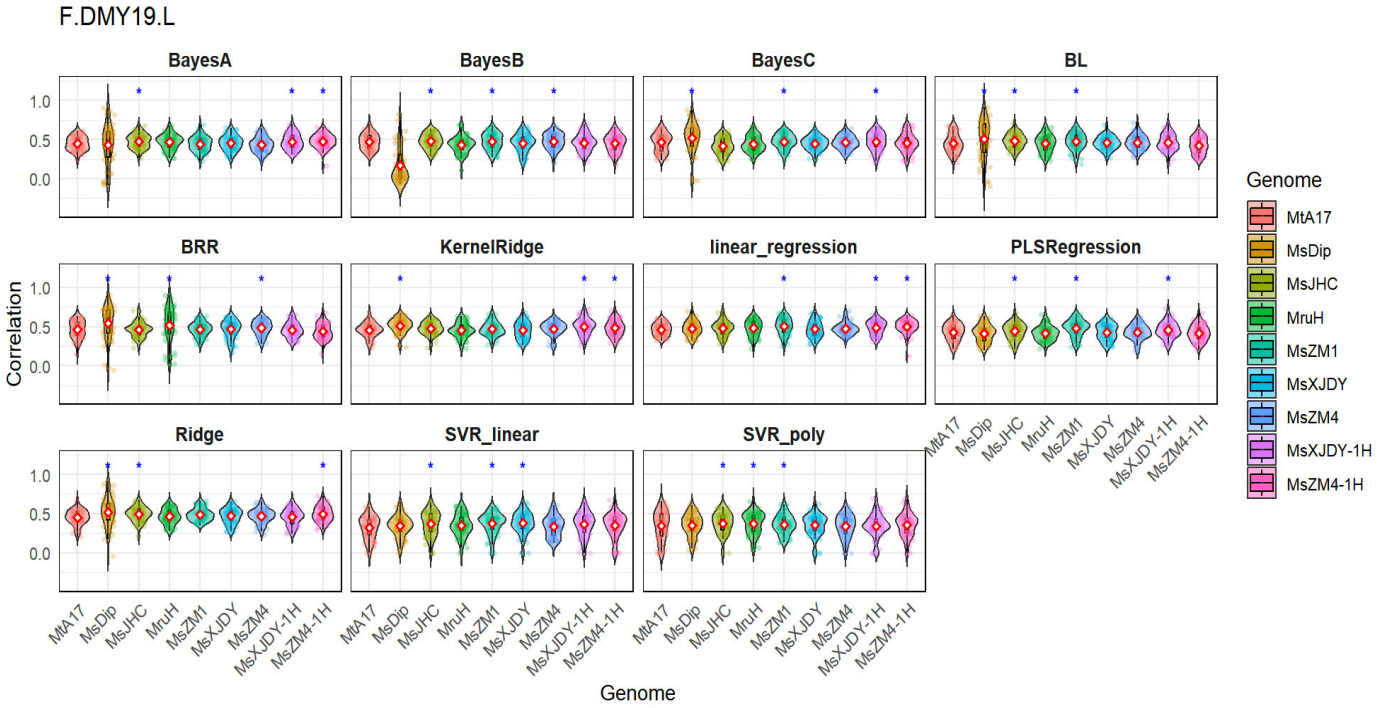

Figure 4. Comparison of prediction accuracy for the F.DMY19.L trait across different reference genomes. The plot displays the prediction accuracy distribution across 9 different reference genomes for 11 models. An asterisk (*) indicates the top three models based on the average prediction accuracy for each model.

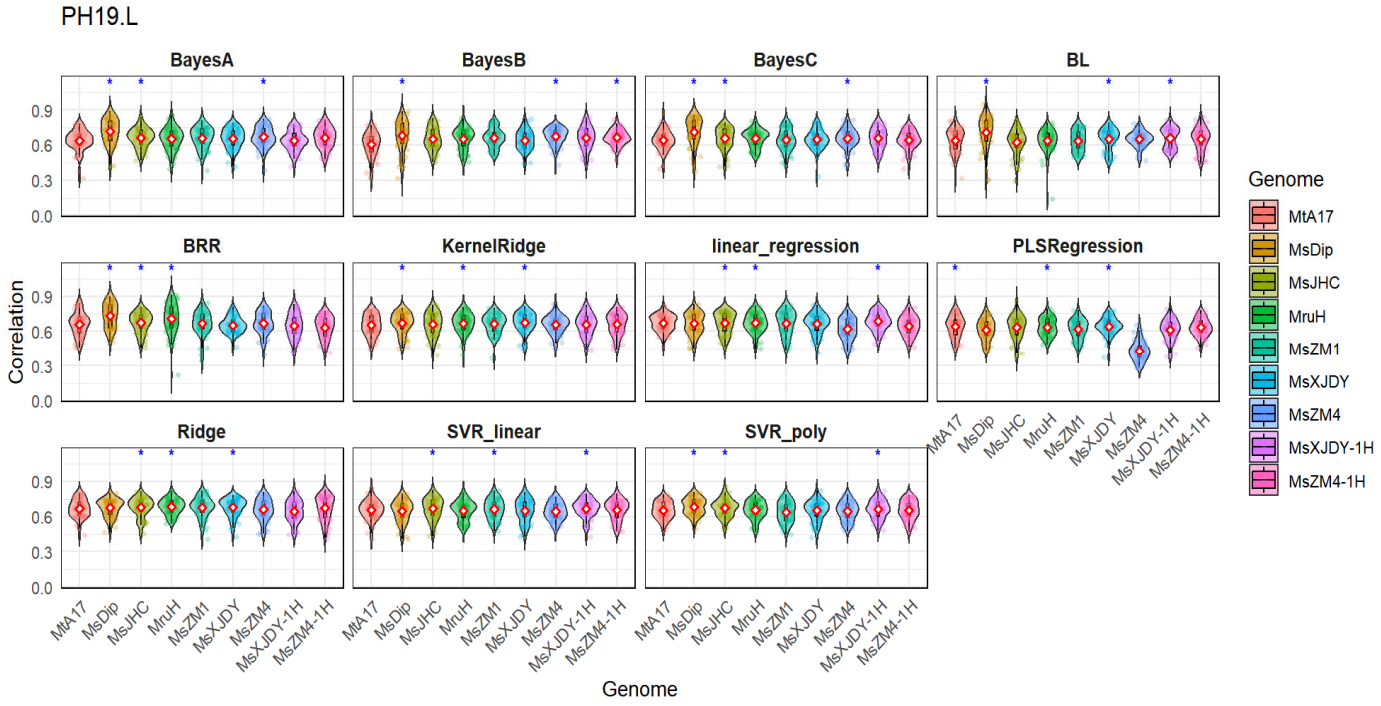

Figure 5. Comparison of prediction accuracy for the PH19.L trait across different reference genomes. The plot displays the prediction accuracy distribution across 9 different reference genomes for 11 models. An asterisk (*) indicates the top three models based on the average prediction accuracy for each model.

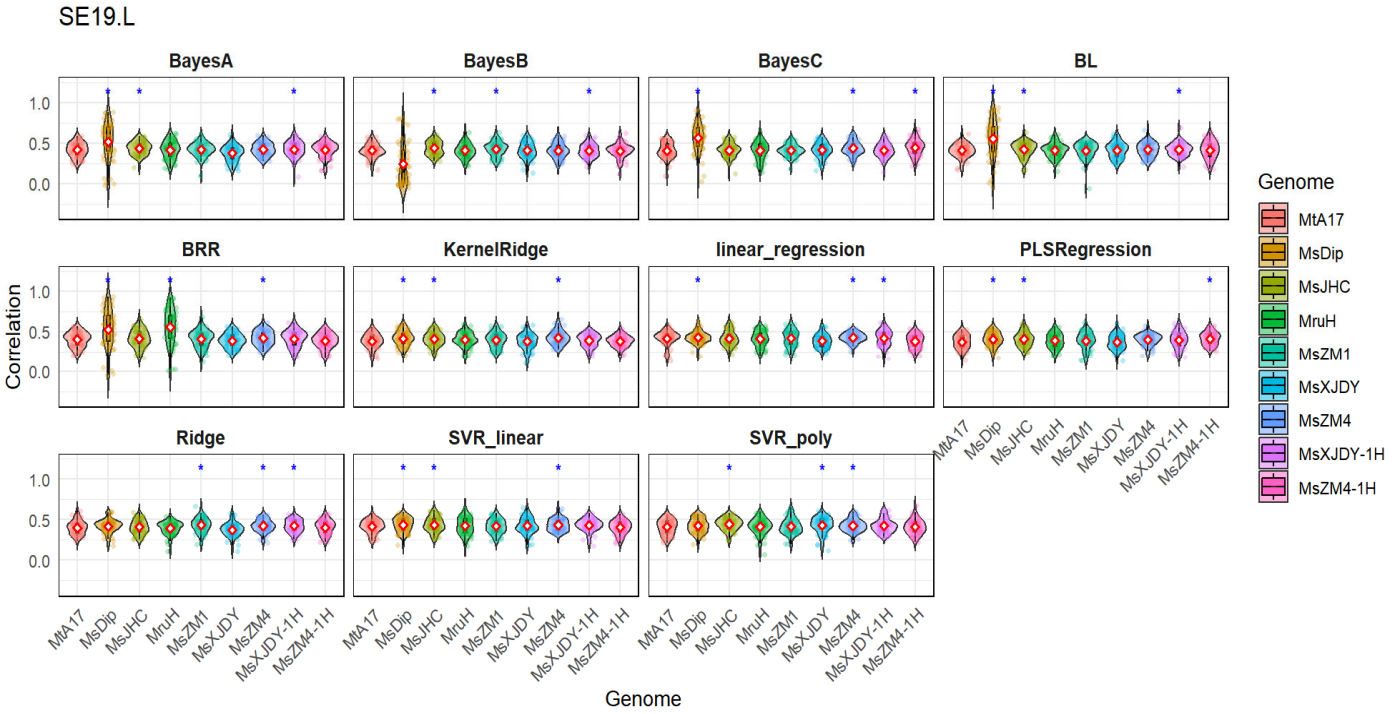

Figure 6. Comparison of prediction accuracy for the SE19.L trait across different reference genomes. The plot displays the prediction accuracy distribution across 9 different reference genomes for 11 models. An asterisk (*) indicates the top three models based on the average prediction accuracy for each model.

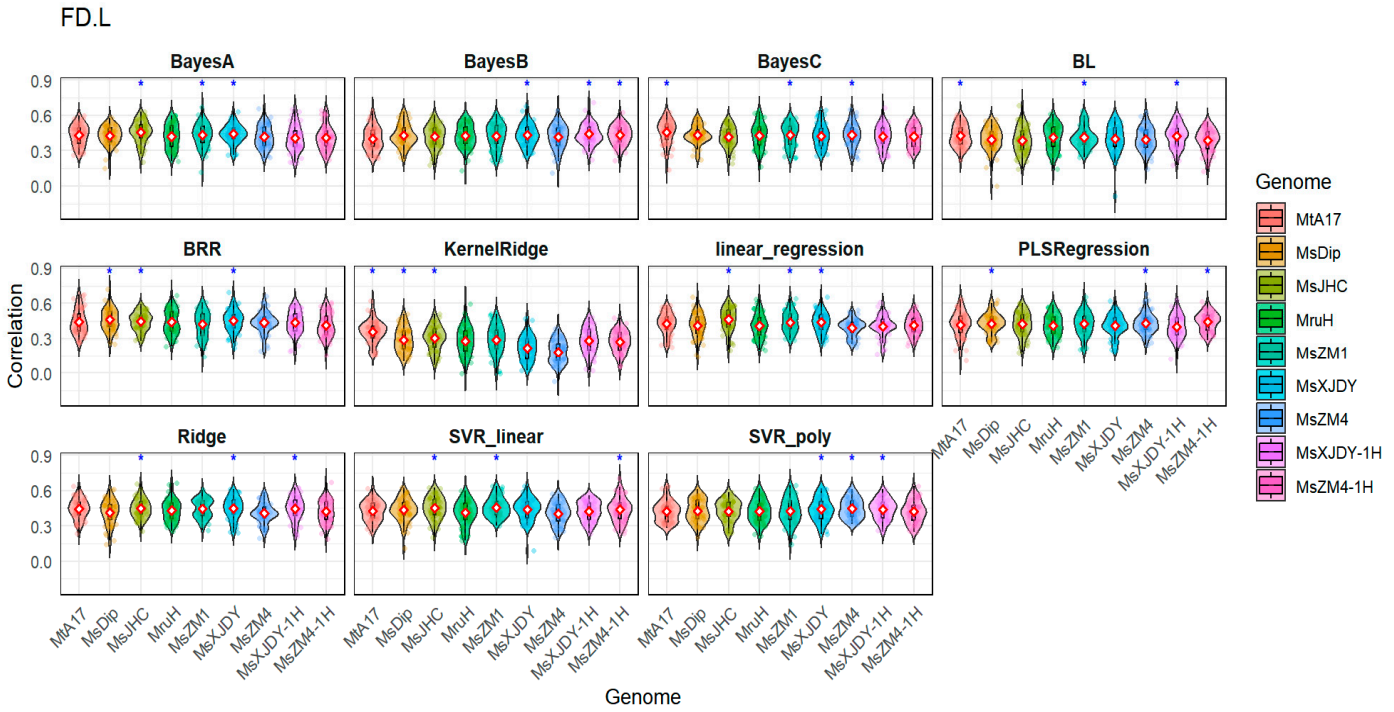

Figure 7. Comparison of prediction accuracy for the FD.L trait across different reference genomes. The plot displays the prediction accuracy distribution across 9 different reference genomes for 11 models. An asterisk (*) indicates the top three models based on the average prediction accuracy for each model.

In the prediction of the D19.L trait, it can be observed that six out of the eleven models showed higher prediction accuracy under the MsDip reference genome. Most models exhibited relatively consistent prediction accuracy across all reference genomes, with correlation values concentrated between 0.6 and 0.7. MsDip performed well in most models, particularly in Bayesian models, where it showed the best performance. Next, MsJHC demonstrated good predictive performance across five models, including BayesA, BayesC, and machine learning models. MsZM1 showed high prediction accuracy in four models: BL, BRR, KernelRidge, and SVR_poly. Additionally, in some models, such as linear_regression and SVR, the monoploid prediction accuracy was also relatively high. This may be related to the correct pairing ratio, which is consistent with the results of the correct pairing rate. However, it is worth noting that the prediction accuracy of MsXJDY and MsZM4 was relatively low, likely due to the complexity of their genomes and the more fluctuating distribution of their correct pairing rates. In contrast, the prediction accuracy for the F.DMY19.L trait was more variable, with significant differences in model performance across different reference genomes. However, it was consistent that the MsDip and MsJHC genomes yielded excellent prediction results, and in some models, the haploid also demonstrated high prediction accuracy. For the PH19.L trait, the prediction results showed consistent correlations across the reference genomes and models, with Bayesian and support vector machine models performing particularly well across all genomes, maintaining correlations above 0.65. MsDip exhibited strong prediction performance across all seven models. In predicting the SE19.L trait, accuracy was generally lower than for other traits, especially in the MtA17 and MsXJDY genomes. Nevertheless, MsDip demonstrated solid prediction performance in eight models. Finally, for the FD.L trait, MsJHC and MsDip exhibited high prediction accuracy in most models. It is worth noting, however, that some reference genomes also showed relatively high prediction accuracy in specific models and traits. This suggests that, although predicting this trait is challenging, certain combinations of reference genomes and models can still provide relatively high prediction performance.

Overall, there are some differences in prediction accuracy across traits with different models and reference genomes, but the diploid genome consistently showed higher and more stable prediction accuracy across multiple traits and models. These results suggest that certain models may perform better for specific traits depending on the reference genome, providing important insights for future genomic selection research.

3.4. Comparison of PVE Values Across Different Reference Genomes

This study evaluated the PVE (Phenotypic Variance Explained) values for five phenotypes across different reference genomes, as shown in Figure 8. In genomic selection, PVE is a key metric that quantifies the contribution of genomic markers to the variation in phenotypic traits, representing the proportion of Phenotypic Variance Explained by the markers. The results indicate that, with the exception of the FD.L trait, the MsDip and MsJHC genomes exhibit particularly high PVE values for the other traits, suggesting these two genomes have stronger predictive power for most traits and may be valuable targets for breeding efforts. For the FD.L trait, the PVE values are generally lower compared to other traits, with values ranging from 0.4 to 0.55 across different genomes. The MsXJDY genome shows the highest PVE value for FD.L, indicating its relatively strong explanatory ability for this trait, whereas the MruH genome has the lowest PVE value, suggesting a more limited capacity to account for phenotypic variation in FD.L.

Overall, the MsDip and MruH genomes consistently show higher PVE values across multiple traits, indicating that they may be ideal reference genomes for explaining variation in these traits during genomic selection. In contrast, the MtA17, MsJHC, and MsXJDY reference genomes show lower PVE values for most traits and may not be the optimal reference genomes for these traits.

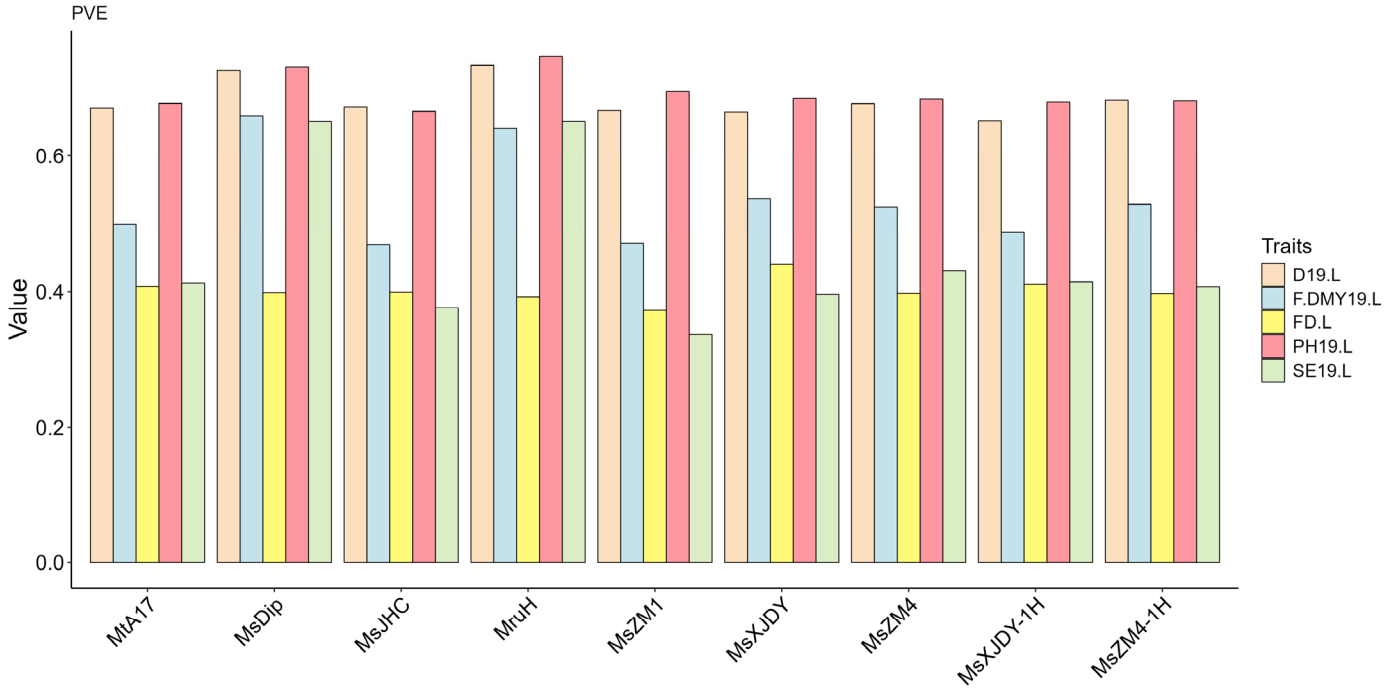

Figure 8. Comparison of Phenotypic Variance Explained (PVE) across different reference genomes for five traits. The x-axis shows the genomes: MtA17, MsDip, MsJHC, MruH, MsZM1, MsXJDY, MsZM4, MsXJDY-1H, and MsZM4-1H. The y-axis represents the PVE values ranging from 0.0 to 0.8. Each bar group corresponds to a genome, while each colored bar within the groups indicates a specific trait: D19.L (light orange), F.DMY19.L (light blue), FD.L (yellow), PH19.L (red), and SE19.L (light green). This figure highlights the capacity of different genomes to explain phenotypic variance for each trait, showing clear distinctions between genomes and their influence across various traits.

3.5. Genomic Selection Prediction Accuracy at Different SNP Densities

From the results of the experiment, MsDip showed overall stability and better performance in genomic selection, indicating that it may serve as an optimal reference genome. Consequently, two traits with high heritability (D19.L and PH19.L) were selected to investigate how varying SNP densities affect the accuracy of genomic selection predictions. Figure 9 shows the genomic selection prediction accuracy for these two high-heritability traits on the MsDip reference genome after SNP thinning. The SNP densities are divided into three groups, 20 k, 50 k, and 100 k, representing SNP selection at intervals of 20 K, 50 K, and 100 K bases, respectively. The x-axis represents different models, including BayesA, BayesB, BayesC, BL (Bayesian Lasso), BRR (Bayesian Ridge Regression), Kernel Ridge, linear regression, PLS regression, Ridge Regression, and Support Vector Regression (SVR) with linear and polynomial kernels. The y-axis indicates the prediction accuracy, used to evaluate the performance of each model under different SNP densities.

For the D19.L trait, overall prediction accuracy significantly improves as SNP density increases. At a 20 K SNP density, most models demonstrate moderate prediction accuracy, ranging from approximately 0.55 to 0.6. When SNP density is reduced to 50 K, the prediction accuracy improves to around 0.6 to 0.7, with models such as BayesA, BayesC, BL, BRR, and Ridge showing the best performance at this density. Further decreasing the SNP density to 100 K results in only slight improvements in some models (e.g., BayesB, PLS regression), suggesting that a 50 K SNP density already provides sufficient genetic information, and additional SNPs do not significantly enhance prediction accuracy. In summary, for the D19.L trait, four models exhibit a trend of decreasing prediction accuracy as density decreases, while five models achieve the highest prediction accuracy at the 50 K density. For the PH19.L trait (Figure 9), the prediction accuracy is generally higher than that for the D19.L trait. A consistent pattern is observed across all machine learning methods and the

BL model: as SNP density decreases, prediction accuracy also declines. Overall, higher SNP densities tend to provide better prediction accuracy; however, certain models can still maintain good prediction accuracy even at lower densities.

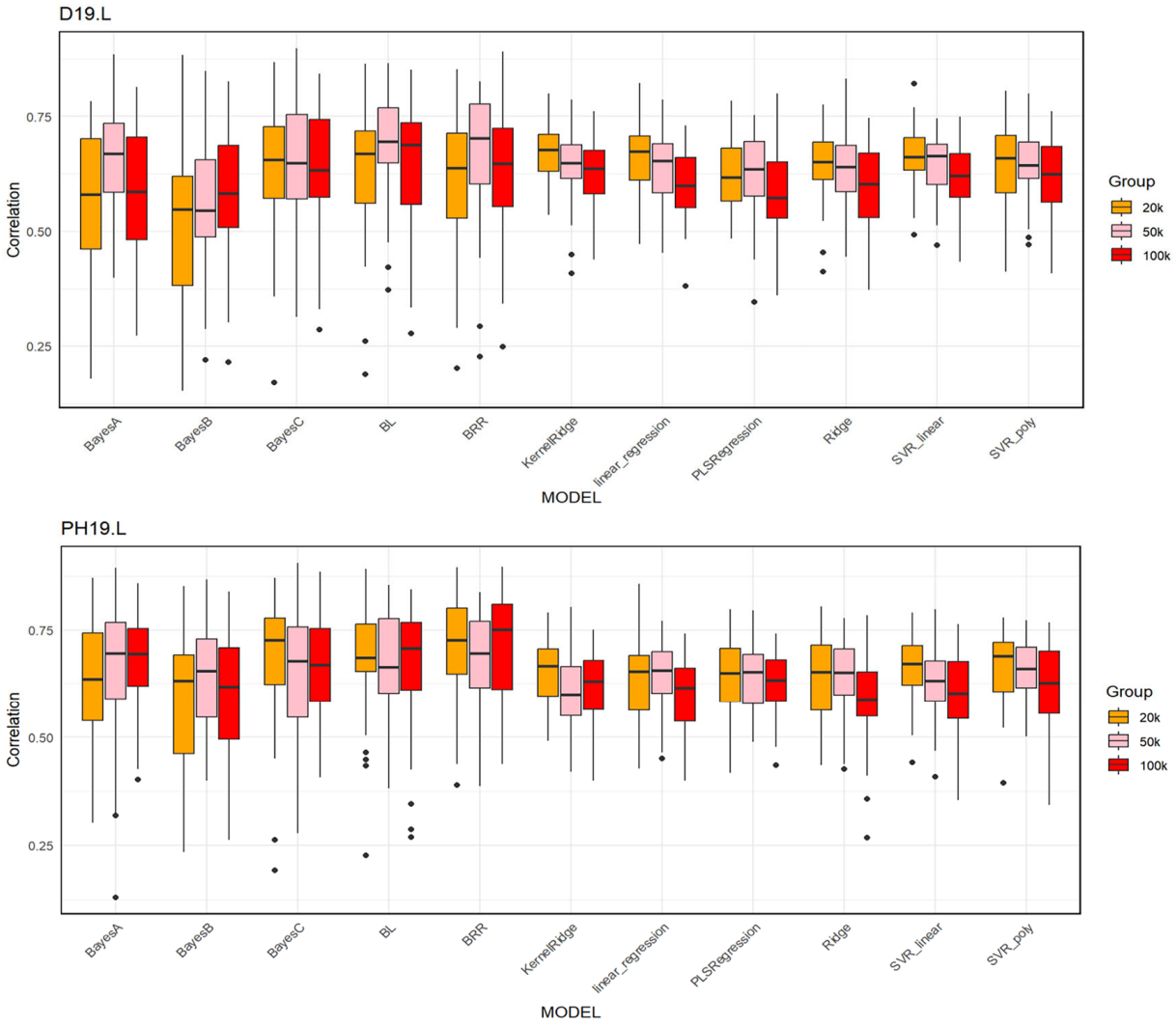

Figure 9. Prediction accuracy of the D19.L and PH19.L traits in MsDip genomes at different SNP densities. The x-axis represents various genomic prediction models, including BayesA, BayesB, BL, Ridge, and others, while the y-axis represents prediction accuracy. Each set of box plots corresponds to three different SNP densities (20 k, 50 k, 100 k), distinguished by different colors.

4. Discussion

In this study, the genomic prediction accuracy of different statistical models was evaluated using phenotypic and genotypic data from mature genetic resources under various reference genomes of alfalfa. Traits such as fall dormancy (D19.L) and plant height (PH19.L) exhibited high heritability across different reference genomes, suggesting a strong genetic basis for these traits. The high heritability values observed in this study are consistent with previous findings in alfalfa breeding, confirming the potential for effective genomic selection on these key traits [36].

By applying multiple genomic prediction models, we were able to assess the relative performance of different approaches. Models such as BayesC, Ridge Regression, and Support Vector Regression (SVR) with a polynomial kernel performed well under higher SNP densities [37]. These models leveraged the genetic information provided by higher marker

density, allowing them to capture complex genetic architectures, which is particularly important for traits like D19.L and PH19.L, which are controlled by multiple loci.

Interestingly, even at reduced SNP densities, models like BayesA and SVR maintained reasonable prediction accuracy, indicating that they can achieve high predictive precision even with fewer markers. This suggests that these models can efficiently utilize the available genetic information, making them suitable for cost-effective breeding programs with limited genotyping resources [38]. This finding is especially significant for resource-constrained breeding programs, where genomic selection can be implemented while balancing costs.

The differences in prediction accuracy across reference genomes and traits underscore the importance of selecting the optimal reference genome for genomic prediction [39]. For instance, the Medicago sativa and diploid genomes provided higher prediction accuracy for high-heritability traits such as D19.L and PH19.L. This highlights the critical role of choosing a reference genome that is genetically aligned with the target population, as the genetic structure of the reference genome profoundly affects the predictive power of genomic selection models [15,16,40]. The variation between traits and genomes suggests that there is no one-size-fits-all solution for genomic prediction [41,42]. Instead, the choice of reference genome and model should be tailored to specific breeding objectives and available resources [2,17]. The performance differences across SNP densities also demonstrate that genomic prediction can be optimized based on breeding goals and resource availability.

Another key aspect is the impact of heritability on model performance. Traits with high heritability, such as D19.L and PH19.L, achieved better prediction accuracy across all models and SNP densities, which aligns with the expectation that high-heritability traits are more suitable for genomic selection [6]. However, for traits with lower heritability, the prediction accuracy of genomic models was more unstable, with a more pronounced decline in predictive accuracy as SNP density decreased [17,29,43–45]. This suggests that breeders should carefully consider the heritability of target traits when designing genomic selection strategies, as traits with lower heritability may require higher SNP densities and more complex models to achieve satisfactory predictive accuracy.

It is important to note that the use of polyploid genomes in plants like alfalfa introduces additional challenges. Compared to diploid genomes, polyploid genomes are much more complex, which makes genomic selection more difficult [46]. This is because polyploid genomes involve multiple copies of genes and intricate genetic interactions, which can lead to variability in prediction accuracy [2]. As a result, selecting the appropriate reference genome and optimizing models remain critical issues in polyploid genome selection. In particular, breeding programs for polyploid plants must carefully consider how to select markers, address the effects of gene copy numbers, and manage the cost of genotyping data.

Furthermore, with the advancements in high-throughput genomics technologies and the accumulation of data, genomic selection models are continuously evolving. Future research could explore more genomic information, such as transcriptomic data, epigenetic information, and metabolomic data, all of which can provide additional genetic insights for breeding [47]. For example, epigenetic modifications may have a significant impact on plant trait expression, particularly under environmental stress conditions [48]. Integrating multi-level omics data can not only improve the accuracy of predictions but also provide more comprehensive theoretical support for molecular breeding, thus promoting the development of precision breeding [49].

Lastly, our study lays a solid foundation for future research to explore more complex genetic architectures, including epistatic interactions and gene-environment interactions. Although this study primarily focused on additive genetic effects, non-additive effects may also play a significant role in trait phenotypic variation, particularly for traits like F-DMY and SE19.L, where prediction accuracy was more variable. Future research could incorporate these non-additive components into genomic prediction models to further improve the predictive accuracy and broaden the application of genomic selection in alfalfa and other forage crops.

5. Conclusions

This study investigates the application of different reference genomes and statistical models in the genomic selection of alfalfa and evaluates their impact on the prediction accuracy of key traits in alfalfa. The results indicate that for all five traits, the prediction accuracy was highest when using the diploid genome. Moreover, in many cases, a positive correlation was observed between SNP density and prediction accuracy: the higher the SNP density, the better the prediction accuracy. There were also differences in the predictive performance of various Bayesian and machine learning models across traits. In conclusion, this study provides valuable insights for precision breeding of alfalfa and other polyploid crops and establishes a theoretical foundation for optimizing genomic selection strategies.

Author Contributions: Conceptualization, Y.S. and X.Z.; methodology, X.Z.; software, X.Z., R.Z., and T.Z.; validation, X.Z., R.Z. and T.Z.; investigation, X.Z.; resources, X.Z., T.Z., C.G. and Y.S.; data curation, X.Z., R.Z. and T.Z.; writing—original draft preparation, X.Z. and Y.S.; writing—review and editing, X.Z., C.G. and Y.S.; visualization, X.Z. and R.Z.; supervision, Y.S.; project administration, Y.S.; funding acquisition, C.G. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Heilongjiang Province (grant number LH2022C050) and the Natural and Science Foundation of China (grant number U21A20182).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: We are grateful to the high-performance computing center at Harbin Normal University for the support of our analysis works. We are also grateful to Manman Li and Shuaixian Li for technical assistance in this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Atanda, S.A.; Govindan, V.; Singh, R.; Robbins, K.R.; Crossa, J.; Bentley, A.R. Sparse testing using genomic prediction improves selection for breeding targets in elite spring wheat. *Theor. Appl. Genet.* **2022**, *135*, 1939–1950. [CrossRef] [PubMed]
2. Alemu, A.; Åstrand, J.; Montesinos-López, O.A.; Isidro, Y.S.J.; Fernández-Gónzalez, J.; Tadesse, W.; Vetukuri, R.R.; Carlsson, A.S.; Ceplitis, A.; Crossa, J.; et al. Genomic selection in plant breeding: Key factors shaping two decades of progress. *Mol. Plant* **2024**, *17*, 552–578. [CrossRef] [PubMed]
3. Werner, C.R.; Gaynor, R.C.; Sargent, D.J.; Lillo, A.; Gorjanc, G.; Hickey, J.M. Genomic selection strategies for clonally propagated crops. *Theor. Appl. Genet.* **2023**, *136*, 74. [CrossRef] [PubMed]
4. Wang, X.; Shi, S.; Ali Khan, M.Y.; Zhang, Z.; Zhang, Y. Improving the accuracy of genomic prediction in dairy cattle using the biologically annotated neural networks framework. *J. Anim. Sci. Biotechnol.* **2024**, *15*, 87. [CrossRef] [PubMed]
5. De Oliveira Celeri, M.; da Costa, W.G.; Nascimento, A.C.C.; Azevedo, C.F.; Cruz, C.D.; Sagae, V.S.; Nascimento, M. Multivariate Adaptive Regression Splines Enhance Genomic Prediction of Non-Additive Traits. *Agronomy* **2024**, *14*, 2234. [CrossRef]
6. Villumsen, T.M.; Su, G.; Guldbrandtsen, B.; Asp, T.; Lund, M.S. Genomic selection in American mink (*Neovison vison*) using a single-step genomic best linear unbiased prediction model for size and quality traits graded on live mink. *J. Anim. Sci.* **2021**, *99*, skab003. [CrossRef]
7. Kaler, A.S.; Purcell, L.C.; Beissinger, T.; Gillman, J.D. Genomic prediction models for traits differing in heritability for soybean, rice, and maize. *BMC Plant Biol.* **2022**, *22*, 87. [CrossRef]
8. Annicchiarico, P.; Nazzicari, N.; Li, X.; Wei, Y.; Pecetti, L.; Brummer, E.C. Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genom.* **2015**, *16*, 1020. [CrossRef]
9. Annicchiarico, P.; Nazzicari, N.; Pecetti, L.; Romani, M.; Ferrari, B.; Wei, Y.; Brummer, E.C. GBS-Based Genomic Selection for Pea Grain Yield under Severe Terminal Drought. *Plant Genome* **2017**, *10*, 2. [CrossRef]
10. Whitmire, C.D.; Vance, J.M.; Rasheed, H.K.; Missaoui, A.; Rasheed, K.M.; Maier, F.W. Using machine learning and feature selection for alfalfa yield prediction. *AI* **2021**, *2*, 71–88. [CrossRef]
11. Zhang, F.; Kang, J.; Long, R.; Li, M.; Sun, Y.; He, F.; Jiang, X.; Yang, C.; Yang, X.; Kong, J.; et al. Application of machine learning to explore the genomic prediction accuracy of fall dormancy in autotetraploid alfalfa. *Hortic. Res.* **2023**, *10*, uhac225. [CrossRef] [PubMed]
12. Kriaridou, C.; Tsairidou, S.; Fraslin, C.; Gorjanc, G.; Looseley, M.E.; Johnston, I.A.; Houston, R.D.; Robledo, D. Evaluation of low-density SNP panels and imputation for cost-effective genomic selection in four aquaculture species. *Front. Genet.* **2023**, *14*, 1194266. [CrossRef] [PubMed]

13. Fu, C.; Hernandez, T.; Zhou, C.; Wang, Z.Y. Alfalfa (*Medicago sativa* L.). *Methods Mol. Biol.* **2015**, *1223*, 213–221. [CrossRef] [PubMed]
14. Chen, H.; Zeng, Y.; Yang, Y.; Huang, L.; Tang, B.; Zhang, H.; Hao, F.; Liu, W.; Li, Y.; Liu, Y.; et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **2020**, *11*, 2494. [CrossRef]
15. Du, C.; Wei, J.; Wang, S.; Jia, Z. Genomic selection using principal component regression. *Heredity* **2018**, *121*, 12–23. [CrossRef]
16. Jia, C.; Zhao, F.; Wang, X.; Han, J.; Zhao, H.; Liu, G.; Wang, Z. Genomic Prediction for 25 Agronomic and Quality Traits in Alfalfa (*Medicago sativa*). *Front. Plant Sci.* **2018**, *9*, 1220. [CrossRef]
17. Medina, C.A.; Kaur, H.; Ray, I.; Yu, L.X. Strategies to Increase Prediction Accuracy in Genomic Selection of Complex Traits in Alfalfa (*Medicago sativa* L.). *Cells* **2021**, *10*, 3372. [CrossRef]
18. Annicchiarico, P.; Nazzicari, N.; Bouizgaren, A.; Hayek, T.; Laouar, M.; Cornacchione, M.; Basigalup, D.; Monterrubio Martin, C.; Brummer, E.C.; Pecetti, L. Alfalfa genomic selection for different stress-prone growing regions. *Plant Genome* **2022**, *15*, e20264. [CrossRef]
19. Pégard, M.; Barre, P.; Delaunay, S.; Surault, F.; Karagić, D.; Milić, D.; Zorić, M.; Ruttink, T.; Julier, B. Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits. *Front. Plant Sci.* **2023**, *14*, 1196134. [CrossRef]
20. Julier, B.; Blugeon, S.; Delaunay, S.; Mappa, G.; Ruttink, T.; Pégard, M.; Barre, P. Optimisation of GBS protocols for efficient genotyping of forage species. In *Exploiting Genetic Diversity of Forages to Fulfil Their Economic and Environmental Roles, Proceedings of the 2021 Meeting of the Fodder Crops and Amenity Grasses Section of EUCARPIA, Freising, Germany, 6–8 September 2021*; Univerzita Palackého v Olomouci: Olomouc, Czech Republic, 2021; pp. 71–74.
21. Li, A.; Liu, A.; Du, X.; Chen, J.Y.; Yin, M.; Hu, H.Y.; Shrestha, N.; Wu, S.D.; Wang, H.Q.; Dou, Q.W.; et al. A chromosome-scale genome assembly of a diploid alfalfa, the progenitor of autotetraploid alfalfa. *Hortic. Res.* **2020**, *7*, 194. [CrossRef]
22. Cui, J.; Lu, Z.; Wang, T.; Chen, G.; Mostafa, S.; Ren, H.; Liu, S.; Fu, C.; Wang, L.; Zhu, Y.; et al. The genome of Medicago polymorpha provides insights into its edibility and nutritional value as a vegetable and forage legume. *Hortic. Res.* **2021**, *8*, 47. [CrossRef] [PubMed]
23. Wang, T.; Ren, L.; Li, C.; Zhang, D.; Zhang, X.; Zhou, G.; Gao, D.; Chen, R.; Chen, Y.; Wang, Z. The genome of a wild Medicago species provides insights into the tolerant mechanisms of legume forage to environmental stress. *BMC Biol.* **2021**, *19*, 96. [CrossRef] [PubMed]
24. Shen, C.; Du, H.; Chen, Z.; Lu, H.; Zhu, F.; Chen, H.; Meng, X.; Liu, Q.; Liu, P.; Zheng, L. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Mol. Plant* **2020**, *13*, 1250–1261. [CrossRef]
25. Long, R.; Zhang, F.; Zhang, Z.; Li, M.; Chen, L.; Wang, X.; Liu, W.; Zhang, T.; Yu, L.X.; He, F.; et al. Genome Assembly of Alfalfa Cultivar Zhongmu-4 and Identification of SNPs Associated with Agronomic Traits. *Genom. Proteom. Bioinform.* **2022**, *20*, 14–28. [CrossRef]
26. Jung, Y.; Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **2022**, *38*, 2404–2413. [CrossRef]
27. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, giab008. [CrossRef]
28. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
29. Liu, J.; Shen, Q.; Bao, H. Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. *PLoS ONE* **2022**, *17*, e0262574. [CrossRef]
30. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
31. Baransel, S.İ.; Ser, G. Evaluation of Beagle Genotype Imputation Method and An Application. *Yuz. Yıl Univ. J. Agric. Sci.* **2017**, *27*, 531–542.
32. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef] [PubMed]
33. Pérez-Rodríguez, P.; de Los Campos, G. Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. *Genetics* **2022**, *222*, iyac112. [CrossRef] [PubMed]
34. Pérez, P.; de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **2014**, *198*, 483–495. [CrossRef] [PubMed]
35. Vojgani, E.; Hölker, A.C.; Mayer, M.; Schön, C.C.; Simianer, H.; Pook, T. Genomic prediction using information across years with epistatic models and dimension reduction via haplotype blocks. *PLoS ONE* **2023**, *18*, e0282288. [CrossRef]
36. Bančič, J.; Ovenden, B.; Gorjanc, G.; Tolhurst, D.J. Genomic selection for genotype performance and stability using information on multiple traits and multiple environments. *Theor. Appl. Genet.* **2023**, *136*, 104. [CrossRef]
37. Burny, C.; Nolte, V.; Dolezal, M.; Schlötterer, C. Highly Parallel Genomic Selection Response in Replicated Drosophila melanogaster Populations with Reduced Genetic Variation. *Genome Biol. Evol.* **2021**, *13*, evab239. [CrossRef]

38. Calus, M.P.; Veerkamp, R.F. Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* **2011**, *43*, 26. [CrossRef]
39. Clark, S.A.; Hickey, J.M.; Daetwyler, H.D.; van der Werf, J.H. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* **2012**, *44*, 4. [CrossRef]
40. Jiang, Y.; Schulthess, A.W.; Rodemann, B.; Ling, J.; Plieske, J.; Kollers, S.; Ebmeyer, E.; Korzun, V.; Argillier, O.; Stiewe, G.; et al. Validating the prediction accuracies of marker-assisted and genomic selection of Fusarium head blight resistance in wheat using an independent sample. *Theor. Appl. Genet.* **2017**, *130*, 471–482. [CrossRef]
41. Riedelsheimer, C.; Czedik-Eysenberg, A.; Grieder, C.; Lisec, J.; Technow, F.; Sulpice, R.; Altmann, T.; Stitt, M.; Willmitzer, L.; Melchinger, A.E. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* **2012**, *44*, 217–220. [CrossRef]
42. Rincent, R.; Laloë, D.; Nicolas, S.; Altmann, T.; Brunel, D.; Revilla, P.; Rodríguez, V.M.; Moreno-Gonzalez, J.; Melchinger, A.; Bauer, E.; et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* **2012**, *192*, 715–728. [CrossRef] [PubMed]
43. Lubanga, N.; Massawe, F.; Mayes, S.; Gorjanc, G.; Bančič, J. Genomic selection strategies to increase genetic gain in tea breeding programs. *Plant Genome* **2023**, *16*, e20282. [CrossRef] [PubMed]
44. Munyengwa, N.; Le Guen, V.; Bille, H.N.; Souza, L.M.; Clément-Demange, A.; Mournet, P.; Masson, A.; Soumahoro, M.; Kouassi, D.; Cros, D. Optimizing imputation of marker data from genotyping-by-sequencing (GBS) for genomic selection in non-model species: Rubber tree (*Hevea brasiliensis*) as a case study. *Genomics* **2021**, *113*, 655–668. [CrossRef] [PubMed]
45. Parveen, R.; Kumar, M.; Swapnil; Singh, D.; Shahani, M.; Imam, Z.; Sahoo, J.P. Understanding the genomic selection for crop improvement: Current progress and future prospects. *Mol. Genet. Genom.* **2023**, *298*, 813–821. [CrossRef]
46. Mbo Nkoulou, L.F.; Ngalle, H.B.; Cros, D.; Adje, C.O.A.; Fassinou, N.V.H.; Bell, J.; Achigan-Dako, E.G. Perspective for genomic-enabled prediction against black sigatoka disease and drought stress in polyploid species. *Front. Plant Sci.* **2022**, *13*, 953133. [CrossRef]
47. Wang, X.; Liu, Z.; Zhang, F.; Xiao, H.; Cao, S.; Xue, H.; Liu, W.; Su, Y.; Liu, Z.; Zhong, H.; et al. Integrative genomics reveals the polygenic basis of seedlessness in grapevine. *Curr. Biol.* **2024**, *34*, 3763–3777.e3765. [CrossRef]
48. Li, D.; Liu, Q.; Schnable, P.S. TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. *Plant Physiol.* **2021**, *186*, 1800–1811. [CrossRef]
49. Gupta, P.K. Quantitative genetics: Pan-genomes, SVs, and k-mers for GWAS. *Trends. Genet.* **2021**, *37*, 868–871. [CrossRef]