Industrial Crops & Products 243 (2026) 123090

Contents lists available at ScienceDirect

Industrial Crops & Products

journal homepage: www.elsevier.com/locate/indcrop

Construction of genomic prediction models for leaf protein content in *Nicotiana tabacum*

Le Yu [a,b,1], Linjie Guo [a,c,1], Li Liu [a,1], Min Ren [a], Lirui Cheng [a], Lei Liang [a,d], Aiguo Yang [a], Huan Si [a], Changchun Cai [e], Yanjun Zan [a,f,*]

[a] Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao 266000, China
[b] Department of Plant Biology, Swedish University of Agriculture Sciences, Uppsala SE-75007, Sweden
[c] College of Agronomy, Qingdao Agricultural University, Qingdao 266109, China
[d] State Key Laboratory of Maize Bio-breeding, Frontiers Science Center for Molecular Design Breeding, Joint International Research Laboratory of Crop Molecular Breeding, National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100091, China
[e] Tobacco Research Institute of Hubei Province, Wuhan 430030, China
[f] Department of Plant Physiology, Umeå Plant Science Center and Integrated Science Lab, Umeå University, Umeå SE-90187, Sweden

ARTICLE INFO

Keywords:
Nicotiana tabacum
Leaf protein content
Genomic selection
Genome-wide association study
Germplasm

ABSTRACT

With its high soluble protein content, large biomass yield, and ease of cultivation, tobacco leaves show strong potential as a novel protein source for livestock. However, the genetic basis underlying leaf protein content remains poorly understood, necessitating the use of genomic prediction models to screen germplasm resources and accelerate the improvement of this trait in future breeding programs. To address this, we analyzed 2517 tobacco germplasm accessions from the Chinese National Tobacco Germplasm Resource Bank, which represent broad genetic diversity, to investigate the genetic architecture of leaf protein content and construct genomic prediction models. Tobacco leaf protein content exhibited a moderate heritability of 0.16, and association analysis identified a significant peak that explained approximately 1% of the phenotypic variance. We further evaluated the performance of 16 mainstream genomic prediction models using five-fold cross-validation. Among these models, best linear unbiased prediction (rrBLUP) model achieved the highest prediction accuracy (0.87). In addition, rrBLUP required less computational time and resources compared with other models, highlighting its stability and efficiency. Field validation (Longshan County, Hunan Province, 111°37′45″E, 27°30′52″N) confirmed the robustness and accuracy of our genomic selection model. Overall, our results demonstrate that genomic prediction can enable rapid screening of tobacco germplasm resources and substantially enhance the efficiency of developing high-protein varieties.

1. Introduction

Over the past few decades, agriculture has undergone numerous innovations aimed at meeting the increasing demands for food production driven by global population growth (Fatica et al., 2021). To reduce feed costs and the environmental impact of the livestock production systems, agricultural research has focused on utilizing agricultural and industrial by-products (i.e., secondary products obtained after harvesting or processing the main product) as forage without compromising the nutrient intake, digestibility, and animal performance (Grasser et al., 1995; Monteiro et al., 2020). Tobacco (*Nicotiana tabacum* L.) has recently gained attention as a potential protein source for livestock due to its outstanding biomass accumulation ability and remarkable adaptability. Fresh tobacco leaves generally contain more than 15% protein, and tobacco leaf proteins are characterized by high purity, low salt content, a balanced amino acid composition, and a higher utilization rate than casein, making them a promising source of edible leaf protein (Shen et al., 2024). Fatica et al. (Fatica et al., 2021) proved that adding Solaris (International Patent PCT/IB2007/053412) (Grisan et al., 2016), a new variety of tobacco, silage biomass to the feed for growing young cows could be beneficial, highlighting its potential as a valuable forage resource and by-product of this multifunctional energy crop.

Tobacco possesses complex defense mechanisms to resist salt stress

* Corresponding author at: Department of Plant Physiolog, Umeå Plant Science Center and Intergrated Science Lab, Umeå University, Umeå, SE-90187, Sweden.
E-mail address: yanjunzan@umu.se (Y. Zan).
[1] These authors contributed equally

https://doi.org/10.1016/j.indcrop.2026.123090

Available online 24 March 2026

(Sun et al., 2020). Our pilot cultivation experiments on severely saline-alkali land in Dongying (118°40′28″E, 37°26′00″N), China confirmed that tobacco could maintain normal biomass accumulation, producing 15–23 tons of fresh tobacco leaves per ha, with some germplasms reaching up to 20% leaf protein content. Combined with its high biomass and ease of cultivation, these traits make tobacco a promising forage resource. Harnessing the potential of this multifunctional crop may help develop feed management strategies in competitive supply chains, particularly in regions traditionally suited for tobacco cultivation but increasingly threatened by land marginalization, abandonment, and degradation (Fatica et al., 2019, 2021).

In recent years, efforts have been made to increase the content of protein in various plants, but traditional breeding approach faces substantial challenges. The Kjeldahl method (Kjeldahl, 1883) and the Dumas method (Dumas, 1831) are both industry standards for nitrogen determination. However, their high cost and slow throughput limit their utility for large-scale phenotyping, reducing the efficiency of conventional phenotype-based breeding (Singh et al., 2025). As protein content is a polygenic trait lacking major effect genes or reliable molecular markers, marker-assisted selection is largely ineffective (Ricroch et al., 2021). Thus, genomic selection (GS) provides a promising alternative by increasing breeding efficiency, enabling precise trait modification, and shortening breeding cycles. Genomic selection, also known as genomic prediction, is an extension of marker-assisted selection to the entire genome (Meuwissen et al., 2001). It leverages genome-wide molecular markers and phenotypic data from reference populations to construct models that estimate genomic breeding values, which can then be applied to predict the performance of offspring (Crossa et al., 2017). GS was initially developed for livestock improvement, and later successfully implemented in major crops such as maize (Peixoto et al., 2024) and rice (Biswas et al., 2023), significantly improved the efficiency of genetic breeding. In summary, GS enables accurate prediction of protein content without direct phenotypic measurement, making it a powerful tool for screening the seedbank and developing high-protein varieties.

GS has only recently been applied in tobacco breeding research, but its potential for genetic improvement is becoming increasingly evident. Carvalho et al. (Carvalho et al., 2022) evaluated 72 hybrid combinations derived from 13 inbred lines to explore the predictive performance of GS in the absence of phenotypic data, demonstrating that the interaction between the genotype and the environment significantly affected influences the prediction efficiency of GS. Tong et al. (Tong et al., 2021) applied GS to recombinant inbred line (RIL) populations of tobacco, testing the effects of marker density, population size, and training-to-testing ratios on predictive performance, with the Bayes B model achieving the highest accuracy. These studies provide critical foundations for applying GS in screening GenBank for elite germplasm, tobacco improvement and offer valuable insights for accelerating the development of high-protein varieties.

In this study, we analyzed 2517 tobacco germplasms with broad genetic diversity to systematically assess genetic variation in leaf protein content. We further evaluated the predictive performance of 16 GS models using five-fold cross-validation, providing a comprehensive benchmark in plant genomics. rrBLUP (ridge regression best linear unbiased prediction) model was identified as the best prediction model on tobacco protein content with a superior behavior on computational efficiency and stability. Our results offer a practical and powerful tool for implementing genomic selection in tobacco protein content and provide both theoretical and methodological foundations for improving tobacco protein content through genomic selection.

2. Materials and methods

2.1. Experimental materials and field design

A total of 2517 tobacco germplasms with broad genetic diversity were selected, including key varieties approved in the past 30 years, genetic materials with significant breeding value as well as commercial varieties. Field trials were conducted in Longshan County, Hunan Province (111°37′45″E, 27°30′52″N) in 2022. The field trials adopted a completely randomized block design with two replicates. The plot area was 22.0 m^2 (row spacing: 1.1 m, plant spacing: 0.6 m, row length: 10.0 m). Each plot was planted with two rows, with one plant per hole, totaling more than 20 plants. The trials were sown on 27 February, and the plants were harvested in three batches when they reached the maturity stage in mid-September. Field water and fertilizer management were carried out according to local planting practices.

2.2. Phenotype identification

To ensure phenotypic consistency across accessions with varying maturation rates, all middle leaves were harvested at the physiological maturity stage of each individual accession—defined by visible yellowing of the leaf surface rather than at a fixed calendar date, allowing each accession to reach its developmental endpoint independently. The 2517 harvested middle leaves were cured in bulk curing barn. Temperature and humidity monitor were used to control temperature and humidity until all leaves turned yellow. After that, preliminary processing was carried out to the cured leaves. The leaves were cut into strands of the same width using rotary cutter to ensure consistency in phenotypic determination. Subsequently, in the Quality and Safety Research Center of the Chinese Academy of Agricultural Sciences Tobacco Research Institute, protein extraction was conducted using the TCA-acetone precipitation method (Granier, 1988), and the content was determined using the BCA method (Smith et al., 1985). The sample protein concentration was calculated based on the standard curve of bovine serum albumin (BSA). During the measurement process, the experimental conditions were strictly controlled to ensure consistency across all samples. The obtained phenotypic data of protein content was used for subsequent genetic analysis and genome selection research.

2.3. Genotype data processing

The genotyping of 2517 germplasms was performed using the genotyping-by-sequencing (GBS) technique (Elshire et al., 2011). First, DNA was extracted by the CTAB method. After being digested by *Nla*III and *Mse*I enzymes, the products were purified using AMPure XP and a sequencing library was constructed. Finally, high-throughput sequencing was completed on the Illumina NovaSeq platform. Raw data were quality-controlled by fastp (Chen et al., 2018) to remove low-quality reads and then aligned to the ZY300 reference genome using BWA (v.0.7.17) (Li and Durbin, 2009). Variations were detected using SAMtools mpileup (v.1.9) (Li et al., 2009). High-quality single nucleotide polymorphism (SNP) loci that were derived by the filter criteria of an average sequencing depth > 2, minor allele frequency (MAF) > 0.03, and individual site missing rates < 0.5 (Poland et al., 2012) (Lu et al., 2013) using BCFtools (v.1.17–50-ga8249495) (Narasimhan et al., 2016). To enhance genotype integrity, missing genotypes were imputed using the Beagle software (Browning and Browning, 2016), and a total of 95,289 high-quality SNPs were obtained.

2.4. Statistical analysis

2.4.1. Kinship Heritability estimation

Using the genomic best linear unbiased prediction (GBLUP) model, the narrow-sense heritability (h^2) (Visscher et al., 2008) of tobacco leaf protein content was calculated through variance component analysis. The calculation formula is as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

h^2 represents the narrow-sense heritability, σ_g^2 represents the genetic

variance, and σ_e^2 represents the residual variance. These parameters were estimated using the restricted maximum likelihood (REML) method to calculate the heritability.

2.4.2. Genome-wide association analysis

Kinship matrix was generated using GCTA (Yang et al., 2011) with –make-grm function. Genome-wide association study (GWAS) was conducted using a linear mixed model implemented in the mlma module of GCTA (Yang et al., 2011). Gene-based association test was performed using the mBAT-combo method (Li et al., 2023) implemented in GCTA, which aggregates SNP-level association statistics within each annotated gene while accounting for linkage disequilibrium (LD) among markers.

2.4.3. Construction of the genomic selection model

Using the 95,289 filtered SNPs as genotype and tobacco leaf protein content as the phenotypic value, genomic selection analysis was carried out. 16 mainstream GS models were evaluated. These models can be roughly classified into three categories. The first category is linear models, including ridge regression (RR) (Whittaker et al., 2000), least absolute shrinkage and selection operator (LASSO) (Usai et al., 2009), and elastic net (EN) (Zou and Hastie, 2005). These models are particularly useful when the genetic structure is sparse, as they can select relevant markers and exclude unimportant ones. The second category is mixed linear models, including genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008), ridge regression best linear unbiased prediction (rrBLUP) (Meuwissen et al., 2001), and non-additive models based on kernel functions such as reproducing kernel Hilbert space (RKHS) (Gianola and van Kaam, 2008) and multi-kernel reproducing kernel Hilbert space (MKRKS) (De Los Campos et al., 2010). Bayesian additive models such as Bayesian A (BA) (Meuwissen et al., 2001), Bayesian B (BB) (Habier et al., 2011), Bayesian C (BC) (George and McCulloch, 1993), Bayesian ridge regression (BRR) (De Los Campos et al., 2013), and Bayesian LASSO (BL) (Park and Casella, 2008). These models make various assumptions about the genetic architecture of traits and are suitable for traits regulated by many minor-effect genes, major-effect genes, and combinations thereof. The third category is nonlinear machine learning models, including deep neural network genomic prediction (DNNGP) (Wang et al., 2023), gradient boosting machine (GBM) (Li et al., 2018), random forest (RF) (Breiman, 2001), and support vector machine (SVM) (González-Recio et al., 2014; Maenhout et al., 2007). These models can capture complex non-additive effects and perform well for traits dominated by genetic interactions.

2.4.4. Cross-validation and model evaluation

To evaluate the predictive performance of the 16 models, 5-fold cross-validation (k = 5) was employed. The population was randomly divided into five equal parts. 20% of the individuals were treated as the test set (containing only genotype data), and the remaining 80% of individuals were used as the training set (with both genotype and phenotype data). The GS model was constructed using the training set, and the phenotypes of the test set were predicted (Zhang et al., 2019). The predictive performance of the models was evaluated using the Pearson correlation coefficient, which is the correlation between the predicted breeding values and the true breeding values. The optimal GS model was selected based on the computational efficiency (running time) and resource consumption across different sample sizes.

3. Results and discussion

3.1. Genetic diversity of 2517 germplasms

We conducted genotype-by-sequencing of 2517 tobacco germplasms from the Chinese National Tobacco Germplasm Resource Bank. This data collection encompasses most of the world's important tobacco germplasms, including the main varieties cultivated in China, K326 and Yunyan 87 (Zan et al., 2025). Principal Component Analysis (PCA) (Patterson et al., 2006) showed that based on genetic similarity, this population could be divided into three subgroups: the dark green part represents introduced tobacco varieties, the purple part represents self-developed tobacco varieties in China, and the yellow part represents local varieties and landraces in China (Fig. 1). This indicates that the population structure is mainly influenced by geographical origin and ancestry (local/selected/introduced).

3.2. Leaf protein content distribution and heritability estimation

Phenotypic analysis of the 2517 germplasms showed that protein content was approximately normally distributed, with a range of 2.32–28.22 mg/g, and an average of 11.81 mg/g (Fig. 2). 72% of the samples were concentrated in the range of 8.41–15.20 mg/g. This indicates that there is a wide range of phenotypic variations in leaf protein content among different germplasm resources.

Heritability estimation shows that the narrow-sense heritability is $h^2 = 0.16 \pm 0.03$ (LRT = 367.25, df = 1, $P < 10^{-100}$). Approximately 20% of the phenotypic variation can be explained by genetic factors. A large number of high-protein germplasm resources provide abundant materials for genetic improvement. The moderate heritability and wide phenotypic distribution range of tobacco leaf protein indicates that this trait has the characteristics of genetic improvement but could be influenced by environmental factors.

3.3. Protein content is controlled by minor-polygenes

To further study the genetic basis of the tobacco leaf protein content, genome-wide association study (GWAS) was conducted using phenotype data from 2517 samples and genotype data obtained from simplified genome sequencing. These results indicated that under the threshold of $P \leq 1 \times 10^{-5}$, a single association peak at chromosome 6 48207318 bp was detected (Fig. 3). This peak explained 0.97% of phenotypic variation, indicating most of the genetic variation is unexplained. The variation of protein content appears to be driven by multiple genes with some being detectable using GWAS. This result is consistent with that of sugar beet and wheat, indicating the polygenic nature of protein content trait (Rijken et al., 2025; Marcotuli et al., 2025). We performed a gene-based association analysis for complex traits using summary-level data from GWAS and linkage disequilibrium (LD) data from a reference sample with individual-level genotypes. This analysis identified 4 genes as the most significant candidate ($P < 0.05$). Among them, NtZY06G00661.1 was annotated as serine/threonine-protein phosphatase 5. PP5 family members are known to regulate stress-responsive signaling and protein turnover pathways (Máthé et al., 2019). NtZY06G00662.1, NtZY06G00663.1 and NtZY06G00664.1 are involved in regulation of transcription, DNA-dependent, G-protein, coupled receptor signaling pathway and tachykinin receptor signaling pathway processes, making them a biologically regulator candidate of leaf protein content, respectively. These results suggest that they may indirectly influence nitrogen assimilation or protein biosynthesis in tobacco leaves, providing candidate targets for future functional validation studies. The selection efficiency is lower when the phenotype is more susceptible to environmental noise and less reflective of the underlying genetic differences. Therefore, improvement of leaf protein content might be more effectively achieved by genomic prediction breeding methods. Overall, this finding provides new evidence for a deeper understanding of the genetic basis of protein content.

3.4. Construction of GS models for predicting tobacco leaf protein content

We investigated the prediction accuracy of 16 GS models (Bayes A, Bayes B, Bayes C, BL, BRR, GBLUP, rrBLUP, LASSO, GBM, EN, RR, RKHS, SVM, MKRKHS, RF and DNNGP). These 16 models represent commonly used statistical and machine learning approaches in GS research and

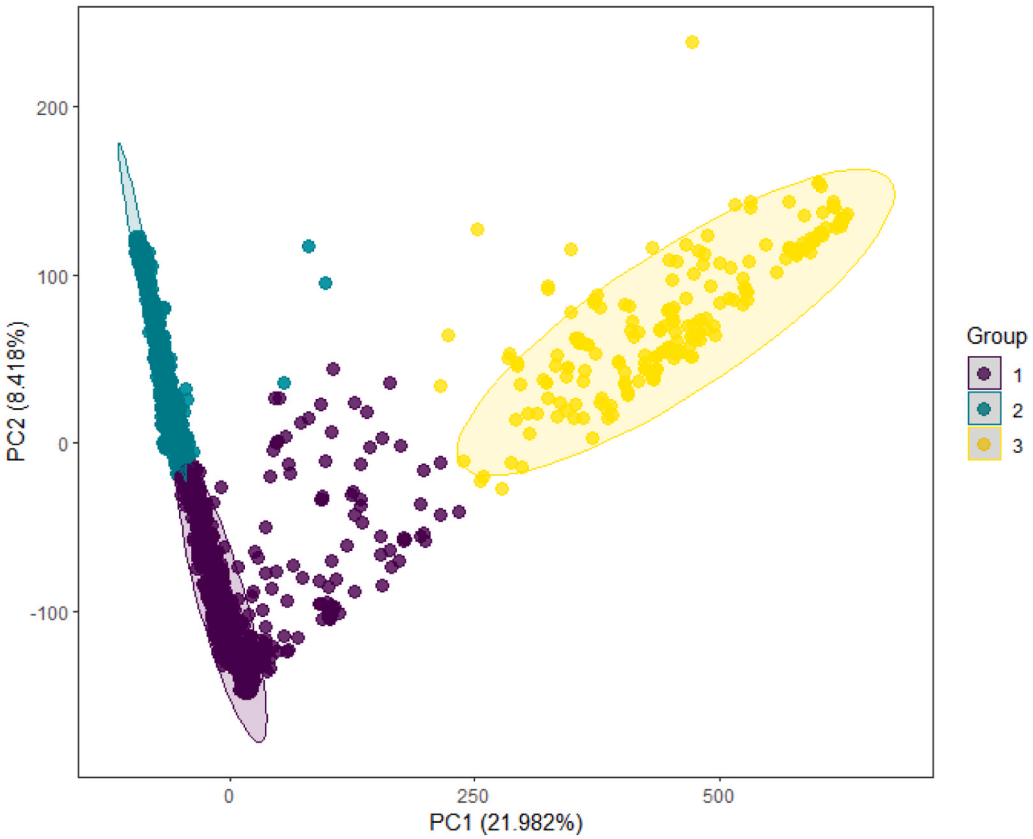

Fig. 1. PCA analysis of tobacco germplasms.

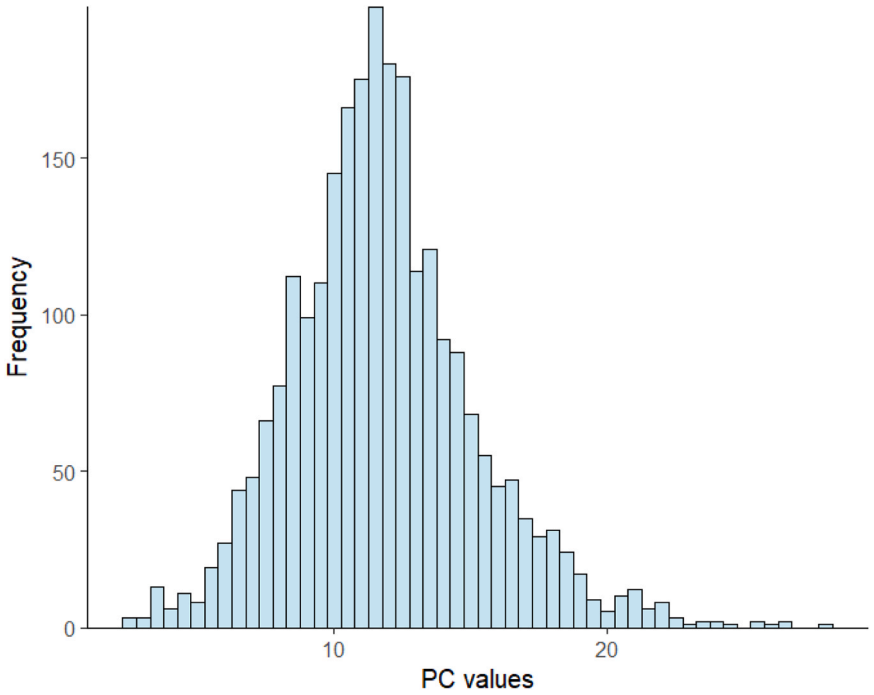

Fig. 2. Phenotypic distribution and heritability estimation of tobacco leaf protein content.

account for different genetic architectures of protein content, enabling a comprehensive exploration of genomic selection for tobacco leaf protein content (Abdollahi-Arpanahi et al., 2020; Heslot et al., 2012). Except for SVM, the prediction accuracy of all models was around 0.4. Among these, rrBLUP model performed the best, with a prediction accuracy of 0.87, followed by the GBM model with a prediction accuracy of 0.81 (Fig. 4).

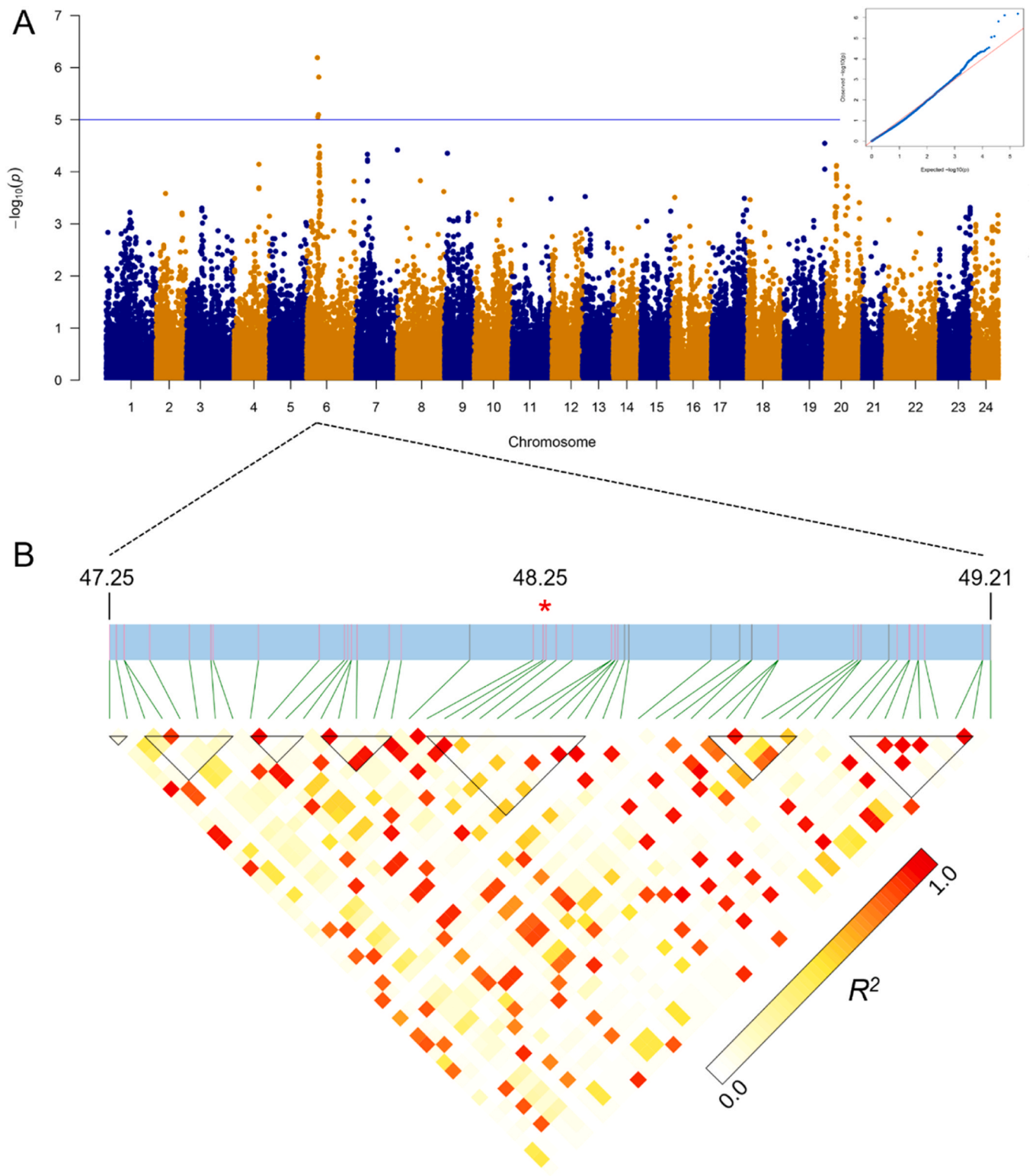

Fig. 3. Association analysis of tobacco leaf protein content. (A) GWAS result across the tobacco genome. (B) Illustration of the LD block around the single peak.

3.5. Comparison of computational speed and resource consumption of 16 genomic selection models

The computational efficiency and resource consumption of the model are important factors affecting large-scale breeding practices. In this study, we compared the computing speed and resource consumption of different GS models when processing large-scale data. The simulation calculation was completed on a laptop equipped with a 12th-generation Intel processor (12 cores, with a main frequency of 2.50 GHz), 16 GB of memory, and running 64-bit Windows 11 system. All analyses were conducted in the R environment (v.4.4.1) (R Core Team, 2024).

We constructed multiple simulated populations with sample sizes of 500, 700, and 1100 using random sampling. To ensure that the genotype data were consistent with the genetic structure characteristics of the

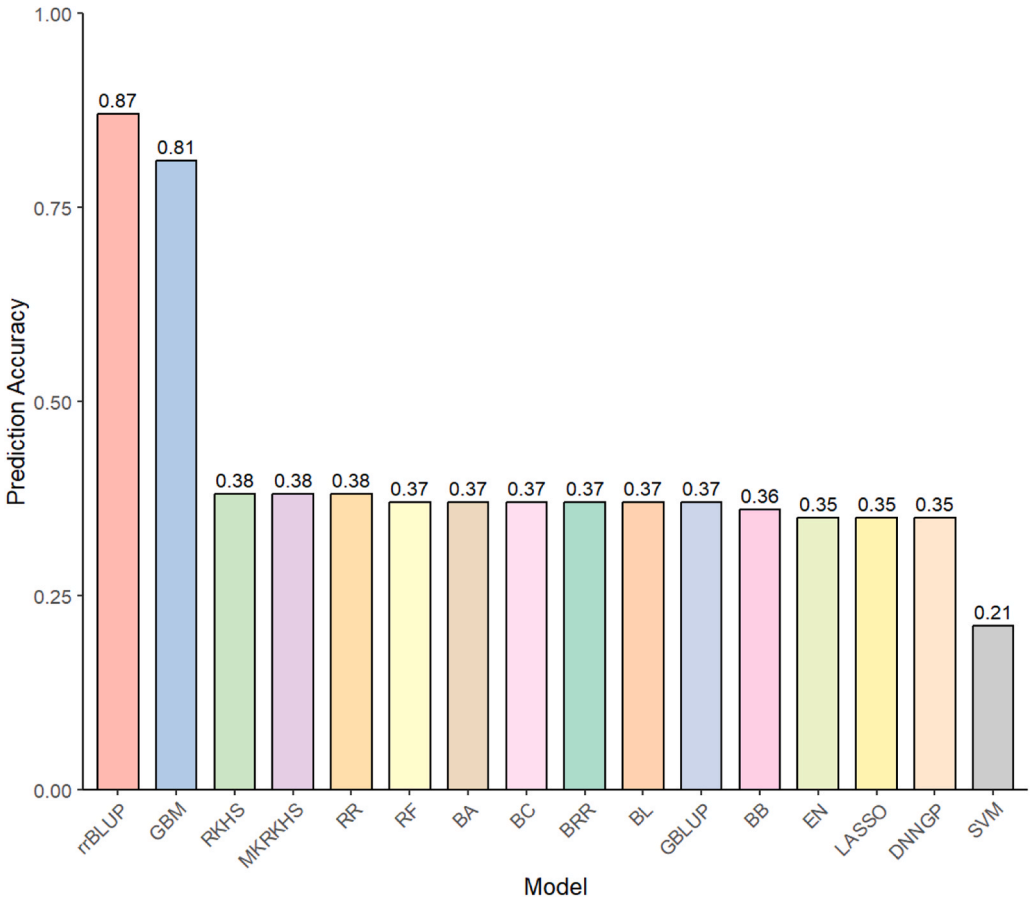

Fig. 4. Prediction accuracy of protein content trait in tobacco.

actual breeding project of tobacco populations, each simulation was completed by a random sampling with replacement from 2517 real tobacco data. Sixteen models were run in each simulated population with different sample sizes. Running time (in seconds) and memory usage of each model were evaluated using five-fold cross-validation. rrBLUP model performed the best in terms of computing speed and low memory consumption under all sample size conditions, especially when the sample size reached 1100 (Fig. 5). Overall, rrBLUP is the best prediction model and is the preferred choice for genomic prediction of protein content.

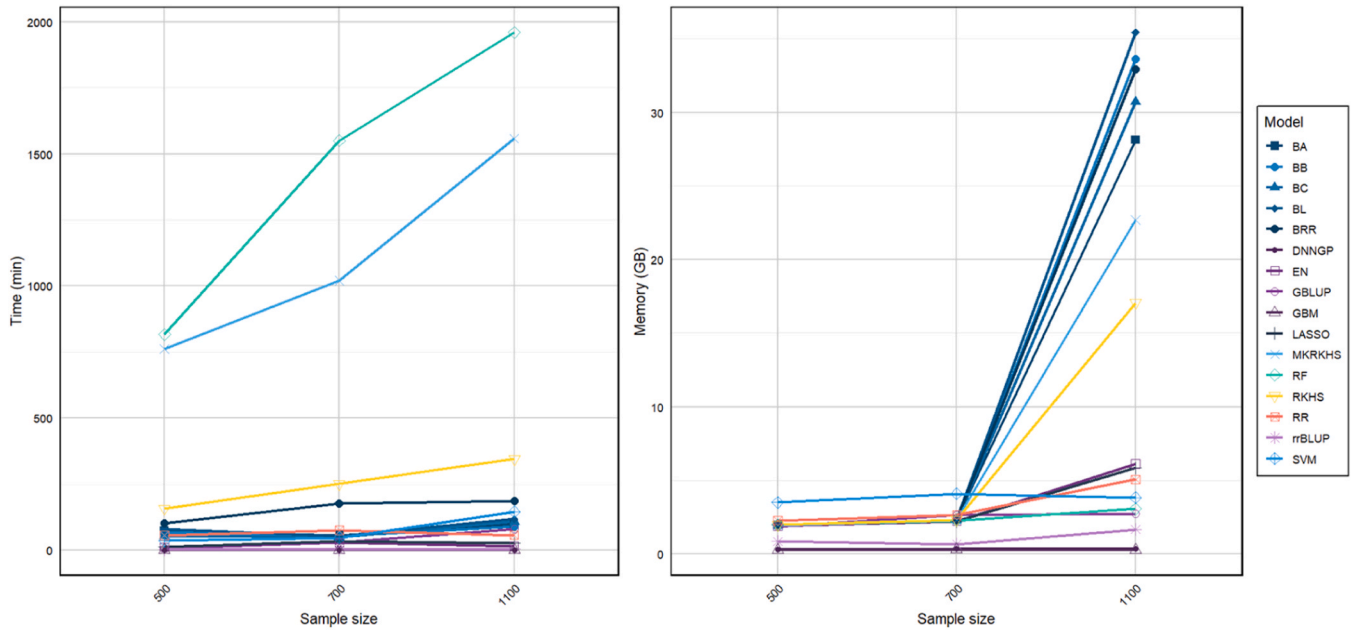

Fig. 5. Comparison of operation speed and computing resources of different models.

3.6. Validation of the prediction model using field test

To validate our prediction model, we predicted protein content of the additional 2983 materials in Chinese National Tobacco Germplasm Resource Bank using the constructed rrBLUP model (Zan et al., 2025). Four germplasms with the lowest protein content (accession no. P1: 2503, P2: 558, P3: 4779 and P4: 1829) and four with highest protein content (accession no. P5: 1532, P6: 1276, P7: 5325 and P8: 3964) were selected and planted in the field in Longshan Country in 2024. Compared with low protein content varieties, high protein content varieties displayed darker leaf color (Fig. 6A). Protein content of eight germplasms was measured and compared with the predicted protein content value. The Pearson correlation between predicted protein content and measured protein content was 0.703, confirming the validity of rrBLUP model for predicting tobacco leaf protein content (Fig. 6B). The insoluble leaf proteins are present in the chloroplasts, accounting for approximately 80% of the total proteins (Pérez-Vila et al., 2022). Our result indicates the potential connection between leaf protein content and leaf color in tobacco. However, selecting for leaf color instead of using a genotypic approach presents several practical challenges. First, leaf color is a later-life trait that is difficult to measure at early developmental stages, necessitating selection in the field. In contrast, genotypic selection can be performed at the seedling stage, significantly shortening breeding cycles and reducing the costs associated with maintaining plants in the field. Second, while the identified associations are significant for extreme accessions, the degree of correlation within a segregating breeding population remains unclear, limiting our confidence in its predictive power for routine selection. While this study did not systematically evaluate genotype-by-environment interactions for protein content, our objective was to develop a predictive model using data from a single site. Despite this limitation, the model—when validated at a second location—successfully identified extreme accessions, demonstrating its potential for screening germplasm. We acknowledge that further multi-site testing is necessary for breeding applications. Nevertheless, this model offers a valuable tool for significantly reducing the number of lines requiring costly phenotyping, thereby streamlining the selection process.

4. Discussion

Unlike morphological traits such as plant height or leaf number that can be assessed visually or with measuring tools in the field, leaf protein content is an invisible quality trait that cannot be determined without laboratory analysis. This inherent phenotyping bottleneck limits the efficiency of conventional phenotype-based breeding and has long been recognized as a major constraint in quality trait improvement across crops (Rasheed et al., 2017; Singh et al., 2025). GS is particularly well-suited to address this challenge: once a prediction model is trained on a reference population with both genotypic and phenotypic data, breeding values can be estimated for candidate lines using genotype data alone (Meuwissen et al., 2001; Crossa et al., 2017). This strategy has demonstrated its effectiveness for selection on quality traits in several major crops. For example, expected genetic gain was 1.4–2.7 times higher with GS than phenotypic selection across quality traits (test weight, 1000-kernel weight, hardness, grain and flour protein, flour yield, sodium dodecyl sulfate sedimentation, Mixograph and Alveograph performance and loaf volume) in wheat panels (Battenfield et al., 2016). In soybean breeding, GBLUP and machine learning demonstrated reliable predictive capability for protein and oil contents (Van der Laan et al., 2025). In terms of tobacco leaf protein content, the rrBLUP model achieved a cross-validation prediction accuracy of 0.87 and an independent field validation correlation of 0.703, demonstrating its practical value for large-scale germplasm screening. These results highlight the advantage of GS for improving complex, laboratory-dependent quality traits in crops.

Despite providing a foundation for the genetic improvement of leaf protein content in tobacco, this study still has limitations. First, this study focused only on markers derived from GBS. The combined use of multi-genomic and epigenomic data would enable more comprehensive capture of genetic variation and potentially further improve prediction accuracy. Second, due to the complex polygenic genetic architecture of leaf protein content, this study did not perform an in-depth dissection of its underlying molecular mechanisms. The single GWAS locus and the accompanying gene-based test collectively account for only a small fraction of the total genetic variance, and the regulatory networks governing nitrogen assimilation, chloroplast protein biosynthesis, and amino acid metabolism in tobacco leaves remain largely unresolved. Future studies could integrate multi-omics data, including transcriptomics, proteomics, and metabolomics, to elucidate the regulatory network controlling leaf protein accumulation on molecular level.

This study demonstrates that GS models have substantial practical value for predicting leaf protein content in tobacco. Prediction accuracy

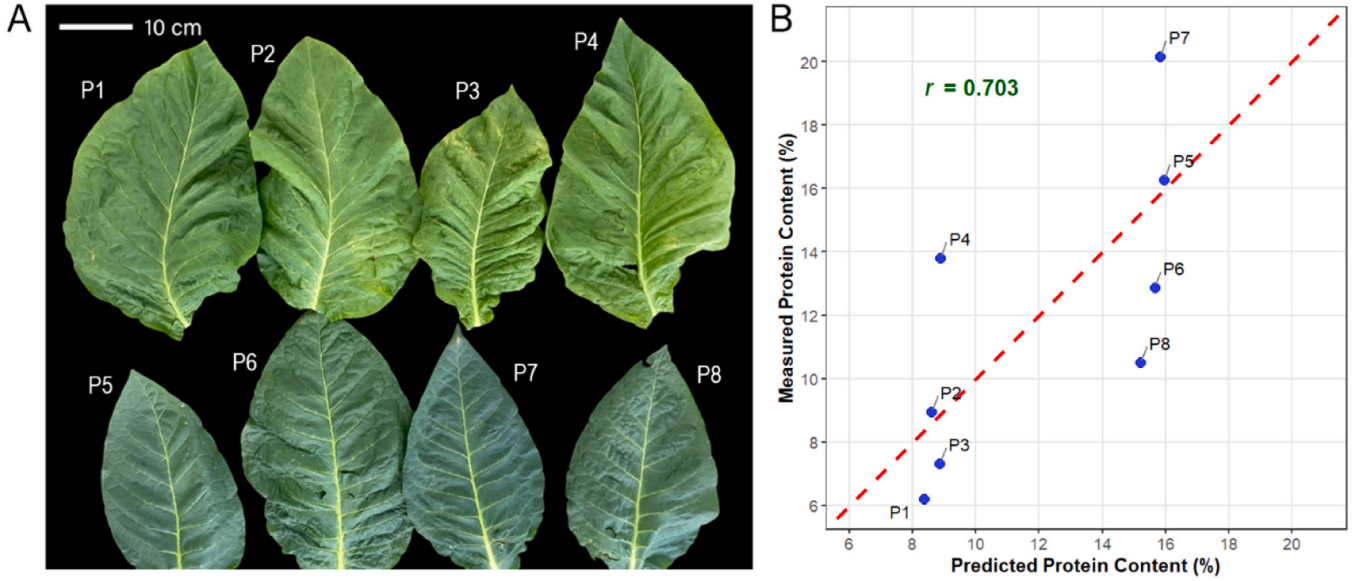

Fig. 6. Validation of the prediction model. (A) Leaf color of four germplasms with low protein content (above: P1-P4) and four with high protein content (below: P5-P8). (B) Prediction accuracy of protein content using rrBLUP model.

could be further enhanced by expanding the training population and increasing marker density through deeper whole-genome sequencing. Combined with tobacco's outstanding biomass yield, adaptability to saline-alkali soils, and the high purity of its leaf proteins, the genomic selection system we established provides critical technological support for the breeding of high-protein tobacco cultivars—facilitating making tobacco a sustainable source of plant-derived protein in the livestock feed industry.

5. Conclusions

In this study, we conducted phenotypic and genetic analyses of the tobacco leaf protein content, revealing extensive phenotypic variation in protein content among different varieties, introduced materials, and local germplasms of tobacco in China. Tobacco leaf protein content was found to be a moderately heritable trait, and there are numerous high-protein genetic materials in the germplasms, indicating substantial potential for genetic improvement of protein content. GWAS analysis identified one locus associated with leaf protein content, explaining 1% of the phenotypic variation. Through the evaluation of prediction accuracy, computational resource consumption, and running time of 16 GS models, together with field validation, a GS technology system for protein content was established based on the rrBLUP model. Our results provide important theoretical and methodological support for revealing the genetic basis of tobacco leaf protein content and GS.

Due to the complex genetic structure of proteins, the genetic mechanism of proteins requires further analysis. Our results provide a candidate site for subsequent functional experiments. Future research should integrate transcriptomics, metabolomics, and other multi-omics approaches to further dissect regulatory network of proteins.

Author contributions

Y.Z and C.C designed and supervised this study. L.Y and L.G implemented the software, Li L, L.C and H.S collected raw data. Lei L, M.R, A. Y, and H.S performed the data analysis. L.Y and Y.Z summarized the results and wrote the manuscript.

CRediT authorship contribution statement

Huan Si: Formal analysis, Data curation. **Changchun Cai:** Conceptualization. **Yanjun Zan:** Writing – review & editing, Supervision, Conceptualization. **Min Ren:** Formal analysis. **Lirui Cheng:** Data curation. **Lei Liang:** Formal analysis. **Aiguo Yang:** Data curation. **Le Yu:** Writing – review & editing, Writing – original draft, Software. **Linjie Guo:** Software. **Li Liu:** Data curation.

Funding

This research was funded by the National Science Foundation of China (32200503), Taishan Young Scholar Program and Distinguished Overseas Young Talents Program from Shandong province (2024HWYQ-079), Key Science and Technology Project from China National Tobacco Corporation (110202101040 JY-17), Jiangsu Tobacco Industrial Co., Ltd (H200407), Hubei Tobacco Research Project (2025KY3CGJ-CYN2A011), China Tobacco Hunan Industry Co., Ltd. Science and Technology Project (KY2025YC0007), Agricultural Science and Technology Innovation Program (ASTIP-TRIC01) from Chinese Academy Agriculture Sciences.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data in the manuscript.

References

Abdollahi-Arpanahi, R., Gianola, D., Peñagaricano, F., 2020. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet. Sel. Evol. 52 (1), 12. https://doi.org/10.1186/s12711-020-00531-z.

Battenfield, S.D., Guzmán, C., Gaynor, R.C., Singh, R.P., Peña, R.J., Dreisigacker, S., Fritz, A.K., Poland, J.A., 2016. Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome 9 (2). https://doi.org/10.3835/plantgenome2016.01.0005.

Biswas, P.S., Ahmed, M.M.E., Afrin, W., Rahman, A., Shalahuddin, A.K.M., Islam, R., Akter, F., Syed, M.A., Sarker, M.R.A., Ifterkharuddaula, K.M., Islam, M.R., 2023. Enhancing genetic gain through the application of genomic selection in developing irrigated rice for the favorable ecosystem in Bangladesh. Front. Genet. 14, *2023*. https://doi.org/10.3389/fgene.2023.1083221.

Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Browning, Brian L., Browning, Sharon R., 2016. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98 (1), 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Carvalho, B.L., Lewis, R., Bruzi, A.T., Pádua, J.M.V., Patto Ramalho, M.A., 2022. Adding genome-wide genotypic information to a tobacco (*Nicotiana tabacum*) breeding programme. Plant Breed. 141 (1), 133–141. https://doi.org/10.1111/pbr.12979.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34 (17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., Varshney, R.K., 2017. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22 (11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011.

De Los Campos, G., Gianola, D., Rosa, G.J.M., Weigel, K.A., Crossa, J., 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92 (4), 295–308. https://doi.org/10.1017/S0016672310000285.

De Los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193 (2), 327–345. https://doi.org/10.1534/genetics.112.143313.

Dumas, J.B., 1831. Procedes de l'analyse organic. Ann. De. Chim. Et. De. Phys. (Ann. Chem. Phys.) 247, 198–213.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos One 6 (5), e19379. https://doi.org/10.1371/journal.pone.0019379.

Fatica, A., Di Lucia, F., Marino, S., Alvino, A., Zuin, M., De Feijter, H., Brandt, B., Tommasini, S., Fantuz, F., Salimei, E., 2019. Study on analytical characteristics of *Nicotiana tabacum* L., cv. Solaris biomass for potential uses in nutrition and biomethane production. Sci. Rep. 9 (1), 16828. https://doi.org/10.1038/s41598-019-53237-8.

Fatica, A., Fantuz, F., Di Lucia, F., Zuin, M., Borrelli, L., Salimei, E., 2021. Ensiled biomass of Solaris tobacco variety used as forage: chemical characteristics and effects on growth, welfare, and follow-up of Holstein heifers. Animal 15 (7), 100235. https://doi.org/10.1016/j.animal.2021.100235.

George, E.I., McCulloch, R.E., 1993. Variable selection via gibbs sampling. J. Am. Stat. Assoc. 88 (423), 881–889. https://doi.org/10.2307/2290777.

Gianola, D., van Kaam, J.B.C.H.M., 2008. Reproducing Kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178 (4), 2289–2303. https://doi.org/10.1534/genetics.107.084285.

González-Recio, O., Rosa, G.J.M., Gianola, D., 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest. Sci. 166, 217–231. https://doi.org/10.1016/j.livsci.2014.05.036.

Granier, F., 1988. Extraction of plant proteins for two-dimensional electrophoresis. Electrophoresis 9 (11), 712–718. https://doi.org/10.1002/elps.1150091106.

Grasser, L., Fadel, J., Garnett, I., DePeters, E., 1995. Quantity and economic importance of nine selected by-products used in California dairy rations. J. Dairy Sci. 78 (4), 962–971. https://doi.org/10.3168/jds.S0022-0302(95)76711-X.

Grisan, S., Polizzotto, R., Raiola, P., Cristiani, S., Ventura, F., Di Lucia, F., Zuin, M., Tommasini, S., Morbidelli, R., Damiani, F., 2016. Alternative use of tobacco as a sustainable crop for seed oil, biofuel, and biomass. Agron. Sustain. Dev. 36 (4), 55. https://doi.org/10.1007/s13593-016-0395-5.

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinf. 12 (1), 186. https://doi.org/10.1186/1471-2105-12-186.

Heslot, N., Yang, H.P., Sorrells, M.E., Jannink, J.L., 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci. 52 (1), 146–160. https://doi.org/10.2135/cropsci2011.06.0297.

Kjeldahl, J., 1883. Neue Methode zur Bestimmung des Stickstoffs in organischen Körpern. Z. F. üR. Anal. Chem. 22 (1), 366–382. https://doi.org/10.1007/BF01338151.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25 (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., 2009. Genome project data processing s: the sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, A., Liu, S., Bakshi, A., Jiang, L., Chen, W., Zheng, Z., Sullivan, P.F., Visscher, P.M., Wray, N.R., Yang, J., Zeng, J., 2023. mBAT-combo: a more powerful test to detect gene-trait associations from GWAS data. Am. J. Hum. Genet. 110 (1), 30–43. https://doi.org/10.1016/j.ajhg.2022.12.006.

Li, F., Zhang, L., Chen, B., Gao, D.Z., Cheng, Y.J., Zhang, X.Y., Yang, Y.Z., Gao, K., Huang, Z.W., Peng, J., 2018. A light gradient boosting machine for remainning useful life estimation of aircraft engines. 21st Int. Conf. Intell. Transp. Syst. (Itsc) 2018, 3562–3567.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., Costich, D.E., 2013. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLOS Genet. 9 (1), e1003215.

Maenhout, S., De Baets, B., Haesaert, G., Van Bockstaele, E., 2007. Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet. 115 (7), 1003–1013. https://doi.org/10.1007/s00122-007-0627-9.

Marcotuli, I., Cabas-Lühmann, P., Caranfa, D., Mores, A., Giove, S.L., Colasuonno, P., Muciaccia, S., Simone, M., Schwember, A.R., Gadaleta, A., 2025. Genome-wide association study for protein and color content in a tetraploid wheat collection. Curr. Plant Biol. 42, 100483.

Máthé, C., Garda, T., Freytag, C., M-Hamvas, M., 2019. The role of serine-threonine protein phosphatase PP2A in plant oxidative stress signaling—facts and hypotheses. Int. J. Mol. Sci. 20 (12), 3028. https://doi.org/10.3390/ijms20123028.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157 (4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Monteiro, L.F. d S., Melo, A.M.P. d, Serrano, M.P., Costa, R.G., Lima, V. d, Medeiros, A.N. d, Lorenzo, J.M., 2020. Suitability of different levels of sunflower cake from biodiesel production as feed ingredient for lamb production. Rev. Bras. De. Zootec. 49, e20190269. https://doi.org/10.37496/rbz4920190269.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., Durbin, R., 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 32 (11), 1749–1751. https://doi.org/10.1093/bioinformatics/btw044.

Park, T., Casella, G., 2008. The Bayesian Lasso. J. Am. Stat. Assoc. 103 (482), 681–686. https://doi.org/10.1198/016214508000000337.

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLOS Genet. 2 (12), e190. https://doi.org/10.1371/journal.pgen.0020190.

Peixoto, M.A., Coelho, I.F., Leach, K.A., Lübberstedt, T., Bhering, L.L., Resende Jr., M.F. R., 2024. Use of simulation to optimize a sweet corn breeding program: implementing genomic selection and doubled haploid technology. G3 Genes|Genomes|Genet. 14 (8), jkae128. https://doi.org/10.1093/g3journal/jkae128.

Pérez-Vila, S., Fenelon, M.A., O'Mahony, J.A., Gómez-Mascaraque, L.G., 2022. Extraction of plant protein from green leaves: biomass composition and processing considerations. Food Hydrocoll. 133, 107902. https://doi.org/10.1016/j.foodhyd.2022.107902.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., Jannink, J.-L., 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5 (3). https://doi.org/10.3835/plantgenome2012.06.0006.

R Core Team (2024). R: A Language and Environment for Statistical Computing. *In R Foundation for Statistical Computing*. ⟨https://www.R-project.org/⟩.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R.K., He, Z., 2017. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol. Plant 10 (8), 1047–1064. https://doi.org/10.1016/j.molp.2017.06.008.

Ricroch, A., Chopra, S., Kuntz, M., 2021. Plant biotechnology: experience and future prospects. Springer Nature.

Rijken, H., Saleem, A., Renedo, J.V., Tossens, A., Bruins, M.E., Trindade, L.M., 2025. Genetic determinants of leaf protein content and extractability in sugar beet: opportunities for leaf biomass valorisation. J. Agric. Food Res. 24, 102498. https://doi.org/10.1016/j.jafr.2025.102498.

Shen, K., Xia, L., Gao, X., Li, C., Sun, P., Liu, Y., Fan, H., Li, X., Han, L., Lu, C., Jiao, K., Xia, C., Wang, Z., Deng, B., Pan, F., Sun, T., 2024. Tobacco as bioenergy and medical plant for biofuels and bioproduction. Heliyon 10 (13). https://doi.org/10.1016/j.heliyon.2024.e33920.

Singh, A.K., Elango, D., Raigne, J., Van der Laan, L., Rairdin, A., Soregaon, C., Singh, A., 2025. Plant-based protein crops and their improvement: current status and future perspectives. Crop Sci. 65 (1), e21389. https://doi.org/10.1002/csc2.21389.

Smith, P.K., Krohn, R.I., Hermanson, G.T., Mallia, A.K., Gartner, F.H., Provenzano, M.D., Fujimoto, E.K., Goeke, N.M., Olson, B.J., Klenk, D.C., 1985. Measurement of protein using bicinchoninic acid. Anal. Biochem. 150 (1), 76–85. https://doi.org/10.1016/0003-2697(85)90442-7.

Sun, H., Sun, X., Wang, H., Ma, X., 2020. Advances in salt tolerance molecular mechanism in tobacco plants. Hereditas 157 (1), 5. https://doi.org/10.1186/s41065-020-00118-0.

Tong, Z., Xiu, Z., Ming, Y., Fang, D., Chen, X., Hu, Y., Zhou, J., He, W., Jiao, F., Zhang, C., Zhao, S., Jin, H., Jian, J., Xiao, B., 2021. Quantitative trait locus mapping and genomic selection of tobacco (*Nicotiana tabacum* L.) based on high-density genetic map. Plant Biotechnol. Rep. 15 (6), 845–854. https://doi.org/10.1007/s11816-021-00713-1.

Usai, M.G., Goddard, M.E., Hayes, B.J., 2009. LASSO with cross-validation for genomic selection. Genet. Res. 91 (6), 427–436. https://doi.org/10.1017/S0016672309990334.

Van der Laan, L., Parmley, K., Saadati, M., Pacin, H.T., Panthulugiri, S., Sarkar, S., Ganapathysubramanian, B., Lorenz, A., Singh, A.K., 2025. Genomic and phenomic prediction for soybean seed yield, protein, and oil. Plant Genome 18 (1), e70002. https://doi.org/10.1002/tpg2.70002.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91 (11), 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Visscher, P.M., Hill, W.G., Wray, N.R., 2008. Heritability in the genomics era — concepts and misconceptions. Nat. Rev. Genet. 9 (4), 255–266. https://doi.org/10.1038/nrg2322.

Wang, K.L., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S., Li, H.H., 2023. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol. Plant 16 (1), 279–293. https://doi.org/10.1016/j.molp.2022.11.004.

Whittaker, J.C., Thompson, R., Denham, M.C., 2000. Marker-assisted selection using ridge regression. Genet. Res. 75 (2), 249–252. https://doi.org/10.1017/S0016672399004462.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88 (1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Zan, Y., Chen, S., Ren, M., Liu, G., Liu, Y., Han, Y., Dong, Y., Zhang, Y., Si, H., Liu, Z., Liu, D., Zhang, X., Tong, Y., Li, Y., Jiang, C., Wen, L., Xiao, Z., Sun, Y., Geng, R., Yang, A., 2025. The genome and GeneBank genomics of allotetraploid Nicotiana tabacum provide insights into genome evolution and complex trait regulation. Nat. Genet. 57 (4), 986–996. https://doi.org/10.1038/s41588-025-02126-0.

Zhang, H., Yin, L., Wang, M., Yuan, X., Liu, X., 2019. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front. Genet. 10, *2019*. https://doi.org/10.3389/fgene.2019.00189.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.