



PREPRINT

Not Peer Reviewed

RESEARCH ARTICLE

Optimizing training set size and marker density for genomic prediction of phytic acid and trypsin inhibitors in soybean (*Glycine max.*)

[version 1]

Norberto Jose Palange^{1,3}, Tonny Obua^{1,2}, Julius Pyton Sserumaga^{id}⁴, Isaac Onziga Dramadri^{1,2}, Ephraim Nuwamanya^{1,5}, Moses Matovu⁴, Mildred Ochwo Ssemakula^{1,2}, Richard Edema^{1,2}, Phinehas Tukamuhabwa^{1,2}

¹Department of Crop Science and Horticulture, Makerere University College of Agricultural and Environmental Sciences, Kampala, Central Region, Uganda

²Makerere Regional Centre for Crop Improvement (MaRCCI), Makerere University College of Agricultural and Environmental Sciences, Kampala, Central Region, Uganda

³Faculty of Food and Agriculture Sciences, Rovuma University, Nampula, Mozambique

⁴National, Livestock Resources Research Institute NaLIRRI), National Agricultural Research Organization, Kampala, --- Select One ---, 000256, Uganda

⁵National, Crops Resources Research Institute (NaCRRI), National Agricultural Research Organization, Kampala, --- Select One ---, 000256, Uganda

v1 First published: 14 Aug 2025, 2:230
<https://doi.org/10.12688/verixiv.1793.1>

Latest published: 14 Aug 2025, 2:230
<https://doi.org/10.12688/verixiv.1793.1>

Abstract

Genomic prediction accuracy is highly influenced by training set size (TSS) and marker density (MD). However, determining the optimal TSS and MD remains trait and model specific. This study assessed the optimum TSS, MD, and statistical models for predicting phytic acid and total trypsin inhibitors (TTI) using a mid-density SNP panel from 308 soybean genotypes. Parametric models, Genomic Best Linear Unbiased Prediction (G-BLUP), Ridge Regression Best Linear Unbiased Prediction (RR-BLUP), Bayes A, B, C, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net, and non-parametric (Random Forest) models were evaluated across training set proportions (0.5–0.9) and marker densities (0.05–1.0). Random Forest consistently outperformed the parametric models in terms of predictive accuracy. The highest prediction for phytic acid was achieved at an 80:20 training-to-test ratio ($R^2 = 0.95$; RPD = 2.81), whereas TTI prediction peaked at the same ratio ($R^2 = 0.97$; RPD =

2.54). Across traits, TSS between 0.5 and 0.9 exhibited minimal differences, and MD above 50% provided negligible gains in model performance. Optimal predictions were TSS of 0.7–0.8, MD of 1.0 for phytic acid, and TSS of 0.5 MD of 0.8 TTI. These findings emphasize the importance of assessing TSS, MD, and prediction models to specific traits for efficient integration into breeding programs.

Keywords

anti-nutritional factors, genomic predictions, marker density, Random Forest, training set size



This article is included in the [Gates Foundation gateway](#).

Corresponding author: Julius Pyton Sserumaga (j.serumaga@gmail.com)

Author roles: **Palange NJ:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Obua T:** Conceptualization, Methodology, Supervision, Visualization, Writing – Review & Editing; **Sserumaga JP:** Data Curation, Formal Analysis, Funding Acquisition, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Review & Editing; **Dramadri IO:** Project Administration, Supervision, Writing – Review & Editing; **Nuwamanya E:** Methodology, Supervision, Writing – Review & Editing; **Matovu M:** Methodology, Supervision, Writing – Review & Editing; **Ssemakula MO:** Project Administration, Supervision, Writing – Review & Editing; **Edema R:** Project Administration, Supervision, Writing – Review & Editing; **Tukamuhabwa P:** Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The Regional Scholarship and Innovation Fund of the Partnership for Skills in Applied Sciences, Engineering, and Technology (RSIF-PASET) and the Carnegie Corporation of New York funded this work. The Genotyping cost was co-funded by the Gates Foundation [grant number OPP1093174] and the Integrated Genotyping Sequence Support (IGSS) project. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2025 Palange NJ *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Palange NJ, Obua T, Sserumaga JP *et al.* **Optimizing training set size and marker density for genomic prediction of phytic acid and trypsin inhibitors in soybean (*Glycine max.*)** [version 1] VeriXiv 2025, 2:230 <https://doi.org/10.12688/verixiv.1793.1>

First published: 14 Aug 2025, 2:230 <https://doi.org/10.12688/verixiv.1793.1>

Introduction

Phytic acid and trypsin inhibitors are anti-nutritional factors (ANFs) with a protective role against pathogens and pests in plants, including soybeans.¹⁻⁴ However, phytic acid and trypsin inhibitors may affect the nutritional value of soybeans in monogastric animals, including humans, by hindering nutrient digestibility and absorption^{3,5} in a dose-dependent manner. Multiple genes control phytic acid and trypsin inhibitors, with each gene harboring a small effect on the overall ANFs expression.⁶ The biosynthetic pathway for Phytic acid and trypsin inhibitors are regulated by signal transduction mediated by phytohormones,⁷ including jasmonic acid (JA), abscisic acid (ABA), ethylene (Eth), salicylic acid (SA), and gibberellins (GA). These phytohormones modulate defense responses by repressing or stimulating cell-to-cell signal transduction.^{8,9} The biosynthesis of phytic acid and trypsin inhibitors is also influenced by environmental factors, adding complexity to their genetic architecture.^{1,10,11}

Breeders often exhibit phenotypic variation for target traits to decide genotype selection. However, regarding the polygenic background of phytic acid¹² and trypsin inhibitors,¹ breeder's ability to make decisions without specialized high-throughput phenotypic data is limited. As phenotyping biochemical traits is often laborious and expensive, selection indices (SI) have been introduced to improve genetic gains. SI combines multiple traits into a single selection value.^{13,14} This method requires careful calibration and a thorough understanding of the trait correlations, economic weights, and genetic architecture.¹³ Therefore, best linear unbiased predictors (BLUPs) have become the best approach for genetic evaluation in SI.¹⁵ With advances in genotyping platforms, new windows have opened for exploring molecular markers associated with specific QTL for important traits in plant and animal breeding.¹⁶ However, marker-assisted selection (MAS) is less effective for polygenic traits such as ANFs.¹⁷ In addition, often available QTL results capture only a limited proportion of the total genetic variance of target traits and require high resources.¹⁸

Genomic selection (GS) offers broader advantages when considering genome-wide marker effects.¹⁹ GS accelerates genetic improvement by reducing time, cost per breeding cycle,^{20,21} and labour-intensive phenotyping.²² The dramatic reduction in genotyping costs in the era of Next-Generation Sequencing (NGS) reinforces the increased utility of GS in crop improvement.²³ The key component of GS is the computation of the individual genomic estimated breeding value (GEBV) in a population to estimate the effects of all genetic loci.¹⁹ To estimate GEBVs, the QTL should be in linkage disequilibrium, with at least one marker distributed throughout the genome.²⁴ Several methods have been proposed to estimate GEBVs, including Genomic-Best Linear Unbiased Prediction (G-BLUP) and Bayesian methods.¹⁹ G-BLUP has been proposed as a convenient method for predicting breeding values and is more suitable for complex, polygenic traits.¹⁹ Ridge Regression Best Linear Unbiased Prediction (RR-BLUP) has been successfully used to generate GEBVs for soybean oil and protein content in soybean.²⁵ Bayesian models can handle traits with complex genetic background.^{21,24,26} The least Absolute Shrinkage and Selection Operator (LASSO) and elastic net (Enet) models have been used to perform variable selection and shrink the regression coefficients simultaneously.^{27,28} Although elastic net offers advantages over LASSO in handling high-dimensional data and when multicollinearity exists among the regressor variables,²⁷ machine learning and deep learning models are reported to be at the forefront of prediction-based data analysis for crop improvement.²⁹ For instance, Random Forest, a machine learning model, has shown superior performance in genomic prediction of soybean compared to other deep learning approaches.²⁹ Machine learning models offer interpretability and facilitate feature selection, allowing the identification of important genetic markers.^{29,30}

The prediction accuracy can be improved by increasing the training set size and marker density.³¹ Optimizing the training set size, composition, genetic relatedness, and marker density is essential for improving the accuracy of genomic predictions^{32,33} by better capturing genetic diversity and increasing statistical power for estimating marker effects.³⁴ Genetic relatedness leads to a higher alignment of genetic backgrounds and improved prediction performance.³⁵

Marker density plays a critical role, as higher densities enable finer mapping of quantitative trait loci (QTL), thus enhancing GEBVs.³⁶ However, continuous increments of training set size and marker density may lead to model over-fitting or inclusion of non-informative markers in the training set,^{21,37} as the benefits of increased marker density depend on the heritability and genetic architecture of specific traits.³⁶ Including genetically diverse individuals in the training set can either improve or reduce prediction accuracy, depending on the genetic complexity of the target trait.³⁸ Therefore, a well-designed training population, accounting for its size, diversity, relatedness, and marker density, is vital for maximizing the efficiency of genomic selection across diverse crop species and traits.³⁹ This study aimed to assess the optimum training set size, marker density, and model for phytic acid and trypsin inhibitor prediction to facilitate genotype selection at an early stage and reduce breeding cycles in soybean breeding schemes.

Material and methods

Plant materials

A set of 308 genotypes from Uganda (117), Taiwan (27), the USA (132), Japan (19), Nigeria (01), and Zimbabwe (12) was obtained from the Makerere University Centre for Soybean Improvement and Development (MAKCSID) program in Uganda.

Field layout and phenotypic traits

To ensure consistent seed multiplication and standardize seed conditions, genotypes were planted in the second growing season of 2023 at Makerere University Agricultural Research Institute Kabanyolo (MUARIK), located in the Central Region of Uganda. The region is at an altitude of 1,180 meters above sea level. The mean annual air temperature is 21.4°C, and the mean annual rainfall is 1,234 mm.⁴⁰ The trial was conducted in a 12 × 26 alpha lattice design with two replications. Independent randomization was conducted using the *agricolae* R package version 4.4.2.⁴¹

Genotyping

The surfaces of young and healthy leaves were cleaned with 70% ethanol and leaf discs were obtained using a manual punch gun. After incubation at 37°C, the samples were referred to SEQART AFRICA hosted at the International Livestock Research Institute in Nairobi for subsequent genotyping by the Diversity Arrays Technology sequencing (DARTseq) complexity reduction method.

Soybean genomic DNA was digested with PstI and MseI enzymes following barcoded ligation with common adapters. Polymerase chain reaction was used for amplification of adapter-ligated fragments, followed by library construction through single-read sequencing runs for 77 bases.⁴² Next-generation sequencing was performed using an Illumina HiSeq 2500 sequencer (Illumina, Inc., Model HiSeq 2500, Rapid Run Mode). Genome profiling was conducted by genotyping by sequencing (GBS) using DARTseqTM technology (Canberra, ACT, Australia). DARTseq marker scoring was performed using the DARTsoft14 scoring pipeline. SilicoDART markers and SNP markers were scored as binary 1 for the presence and 0 for the absence of restriction fragments with marker sequences in the genomic representation of the sample. Chromosomal positions were determined by alignment with the soybean-Wm82-a1-v4 reference genome.^{42–44}

Phytic acid and TTI phenotyping

Phytic acid was extracted using the acidic method,⁴⁵ with minor adjustments to the initial sample volume and refrigeration during centrifugation. For total trypsin inhibitors, a neutral method⁴⁶ was applied, with similar modifications in sample size, volume, and centrifugation conditions. Phenotypic data were generated as described in our earlier publications.^{47,48}

Data analysis

SNP marker quality, outlier detection and data processing

Single nucleotide polymorphism (SNP) data were filtered with a threshold of 95% reproducibility and 95% call rate, 0.05 minor allele frequencies (MAF < 0.05), and imputation through k-nearest neighbor imputation (*knni*) using *snpReady* package in R software version 4.4.2.⁴¹ Duplicates were removed from the filtered SNP markers using the *duplicated* function of *dplyr* package⁴⁹ in the R software version 4.4.2.⁴¹ Of the 18,300 initial SNP markers, 11,805 quality SNP markers were used for further analysis.

The interquartile range (IQR) method was applied to remove outliers from the target variables, including phytic acid and total trypsin inhibitors (TTI), and the distribution and descriptive statistics were calculated using *dplyr*⁵⁰ package in R version 4.4.2.⁴¹ The *cor* function was used to compute The Pearson correlation coefficient between traits was computed using the *Cor* function. To assess whether the variances for phytic acid and TTI were homogenous between the training and test sets, Levene's test was applied using *car*⁵¹ package in R version 4.4.2.⁴¹ at *p*-value = 0.05.

Parametric and non-parametric models

Parametric (linear regression) and non-parametric (machine learning) models have been proposed for the genome-based prediction of genetic values.⁵² Each method differs in the marker effect assumptions.^{24,53} Parametric methods use a linear mixed model to analyze marker additivity for a specific trait. This approach is incapable of predicting phenotypes with an epistatic genetic architecture.⁵² Therefore, we retested the merit of parametric models, including G-BLUP, RR-BLUP, Bayesians, Elastic Net, and LASSO, against non-parametric Random Forest using *rrBLUP*,⁵⁴ *BGLR*,⁵⁵ *glmnet*⁵⁶ and *randomForest*⁵⁷ packages, respectively, in R version 4.4.2.⁴¹ In parametric models, such as G-BLUP, the genomic

relationship matrix is computed assuming that markers contribute equally and additively to the trait variance, without estimating individual marker effects. The general formula of the model can be represented as:

$$y = X\beta + Zm + \epsilon,$$

where β is a $p \times 1$ vector of unknown fixed effects; usually the first element is the population mean, and X is the incidence matrix that relates y to β ; m is a vector of random marker effects; ϵ , the vector of residual errors.⁵²

RR-BLUP is a linear regression with a ridge penalty. The model assumes a normal distribution of marker effects and has identical variance. This model executes genome-wide association analysis based on a mixed model^{24,54}:

$$y = X\beta + Zg + S\tau + \epsilon,$$

where β represents a vector of fixed effects modeling both environmental factors and population structure, g models the genetic background of each line as a random effect with $\text{Var}[g] = K\sigma^2$, τ models the additive SNP effect as a fixed effect, and residual variance is given by $\text{Var}[\epsilon] = I\sigma^2$.

The Bayesian models (A, B, and C) used Bayesian inference and linear modeling in the context of genetic data. However, Bayes A assumes that all markers have independent effects with different variances, whereas Bayes B assumes that most markers have no effect, with only a few markers contributing to the phenotype. The Bayes C method allows for a subset of markers to have non-zero effects using a flexible mixture model approach.²⁴ Based on these methods, for individual i the mathematical model can be written as:

$$y = \mu + \sum_{j=1}^p X_{ij}mj + ei,$$

where $i = 1 \dots n$ individual, $j = 1 \dots p$ marker position or segment, y_i represents the phenotypic value for individual i , m is the $n \times 1$ dimensional general mean vector, X_{ij} represents an element of an incidence matrix for marker j and individual i , m_j is the random effect for marker j , and ei the random residual. The Bayesian model is expressed as follows.⁵²

$$y = \mu + \sum_{j=1}^p X_{ij}mj + e.$$

The elastic net model is treated as the corrected phenotype and can be expressed as $y = Z\gamma + e = \sum_{k=1}^m Z_k\gamma_k + e$; where y represents a vector for the corrected phenotype, Z , matrix of genotype codes for SNPs with -1, 0, and 1; γ is the vector of SNP effects; $e \sim N(0; I\sigma_e^2)$ represents the residual error vector; Z_k the vector of SNP genotypes for n individuals at locus k for $k = 1, \dots, m$. m represents the number of markers n , is the number of individuals, γ_k is the SNP effect for marker k . Ordinary least squares (OLS) estimates parameter γ by minimizing the residual sum of squares⁵⁸:

$$y^{RSS} = \arg \min_{\gamma} \epsilon \Omega \sum_{i=1}^m \left(y_i - \sum_{k=1}^m Z_{ik}\gamma_k \right)^2.$$

The LASSO penalty estimates marker effects by adding an L1-norm to OLS⁵⁸:

$$y^{LASSO} = \arg \min_{\gamma} \epsilon \Omega \left\{ \sum_{i=1}^m \left(y_i - \sum_{k=1}^m Z_{ik}\gamma_k \right)^2 + \lambda \sum_{k=1}^m |\gamma_k| \right\}.$$

Unlike parametric models, non-parametric models capture the complex interactions among the markers. For instance, Random Forest (RF), a machine-learning algorithm, uses regression trees and captures the interactions between markers. The best prediction model was obtained by averaging the tree nodes.⁵⁹ In the RF, prediction is given by the averaging of individual tree predictions as follows: $f_{Brf}(x) = f^{rf}(x) : 1/B \sum_{b=1}^B Tb(x)$ ⁶⁰; where $f^{rf}(x)$ represents the final prediction made by the RF model for an input x ; B is the total number of trees in the RF ensemble; $Tb(x)$ the prediction made by the b -th tree for the input x . Each tree Tb is trained on a bootstrapped subset of the data, and it makes a prediction independent; the $1/B \sum_{b=1}^B$ represents the averaging process for the predictions from all B trees.

In this study, the training set for all parametric and non-parametric models was 80%. Based on the predictions obtained, the Random Forest (non-parametric) model exhibited the best predictions compared to parametric models, including G-BLUP, RR-BLUP, Bayesian, Elastic Net, and LASSO. Therefore, RF was subsequently used for further analysis, including optimization of the training set size and marker density.

Effect of training set size and marker density

Different training set sizes were assessed. The evaluated proportions were 0.5, 0.6, 0.7, 0.8, and 0.9, with 500 decision trees and ten-fold cross-validation for model robustness using the *caret* and *randomForest* packages⁵⁷ in R software. The effect of marker density on prediction accuracy was assessed at 0.05, 0.1, 0.25, 0.5, 0.7, 0.8, and 1.0, using the *reshape2*⁶¹ package in R version 4.4.2.⁴¹

Prediction metrics included the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and ratio of performance to deviation (RPD). The performance metrics were visualized using the *ggplot2* package⁵⁰ in R, version 4.4.2.⁴¹

A 0.95 quantile threshold was applied to the initial set of 11,805 high-quality SNP markers to identify the most variable ones, yielding 590 SNPs. The top 10 SNPs and their associated genotypes were visualized using a heatmap generated using the *heatmap* package in R.⁴¹

To account for the role of environmental factors in prediction accuracy,⁶² variance components were estimated using multivariate linear mixed models (MLMMs) fitted through restricted maximum likelihood (REML). Broad- and narrow-sense heritability and GEBVs were computed using the *lme4*, *sommer*,⁶³ and *randomForest*⁵⁷ packages in R version 4.4.2,⁴¹ respectively.

Results

Descriptive statistics and trait distribution

The mean phytic acid content in the samples was 1.72 mg g^{-1} (SD = 1.20), and total trypsin inhibitors (TTI) had a lowest mean of 0.85 mg g^{-1} (SD = 0.29) (Table 1). A weak positive relationship between phytic acid and TTI ($r^2=0.2$) was detected using Pearson's correlation test. Levene's test confirmed the homogeneity of variances between the training and test sets ($p > 0.05$). The phenotypic distribution using Gaussian kernel density estimates across the training, test, and total sets closely followed a normal distribution (Figure 1).

Prediction accuracy using parametric and non-parametric statistical models

Random Forest, a non-parametric model, showed high prediction accuracies for phytic acid and TTI (Table 2). Among the parametric models tested, LASSO achieved the highest prediction accuracy for phytic acid (85%). In contrast, the prediction accuracy for the TTI was low (approximately 20%). The G-BLUP predictions were approximately 22% and 4% for phytic acid and TTI, respectively. All other models showed prediction accuracies below 20% for phytic acid and nearly zero for TTI. Therefore, further genomic prediction analysis was conducted using only the Random Forest machine learning model.

Training and cross-validation of the genomic prediction

The training and cross-validation results obtained using the 80% training set showed a high predictive accuracy for both traits (Table 2). Phytic acid and TTI had R^2 values of 95% and 97%, respectively, for the test set. The regression curve comparing the predicted and actual values further supported the reliability of the model (Figure 2).

Effect of training set size and marker density

Prediction accuracies obtained using training set proportions between 0.5 and 0.9 showed no substantial differences. However, slight variations were observed between traits within the same training set (Table 3). Conversely, increasing marker density led to a slight improvement in prediction, with minimal differences between traits (Figure 3A and 3B).

Table 1. Descriptive statistics for phytic acid and TTI.

Trait	Set	Mean	Median	SD	Min (mg g^{-1})	Max (mg g^{-1})
Phytic acid	Training	1.72	1.50	1.17	0.02	6.13
	Test	1.89	1.66	1.28	0.20	6.93
	Total	1.75	1.52	1.19	0.01	6.93
TTI	Training	0.85	0.91	0.29	0.06	1.53
	Test	0.86	0.92	0.28	0.01	1.34
	Total	0.85	0.91	0.29	0.01	1.53

SD – standard deviation; Min – minimum; Max – maximum.

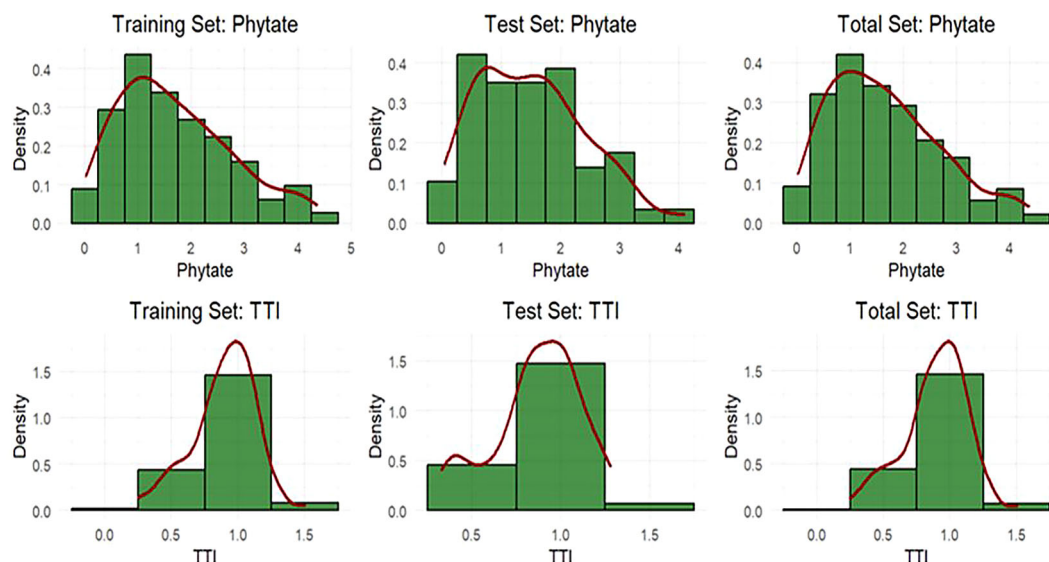


Figure 1. Phenotypic distribution of phytic acid and TTI in training, test, and total sets. The normal distribution supports the suitability of statistical methods assuming normality. Gaussian kernel density estimates for phytic acid and TTI in training, test, and total sets are depicted in red. Distribution of phytic acid shows a single peak at approximately 1.2 mg. g⁻¹ in the training set, two peaks at around 0.8 mg. g⁻¹ and 1.8 mg. g⁻¹ in the test set, and a single peak at approximately 1.2 mg/g in the total set. For TTI, the training set exhibits a peak at around 1.9 mg/g, while the test set has two peaks at approximately 0.3 mg. g⁻¹ and 0.9 mg. g⁻¹. The total set shows a single peak at around 1.0 mg. g⁻¹.

Table 2. Prediction statistics for phytic acid and TTI using 80:20 training-to-test set ratio.

Sets	Trait	Method	R ² _{cv}	RMSE	MAE	MSE	ntree	RPD	Biases
Training	Phytic acid	RF	0.96	0.43	0.33	1.01	500	2.74	0.025
	TTI	RF	0.98	0.14	0.12	0.10	500	2.54	0.005
Testing	Phytic acid	RF	0.95	0.42	0.32	1.21	500	2.81	0.001
	TTI	RF	0.97	0.11	0.09	0.09	500	2.54	0.002

TTI – total trypsin inhibitors; rf – random forest; R² – coefficient of determination; RMSE – root mean square error; MAE – mean absolute error; MSE – means square error; *ntree* – number of trees; RPD – ratio of performance to deviation.

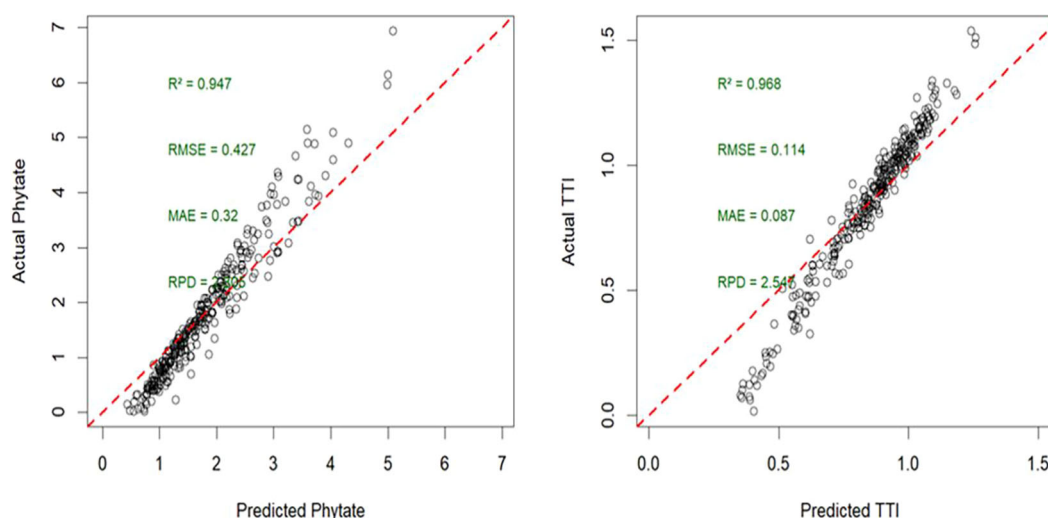
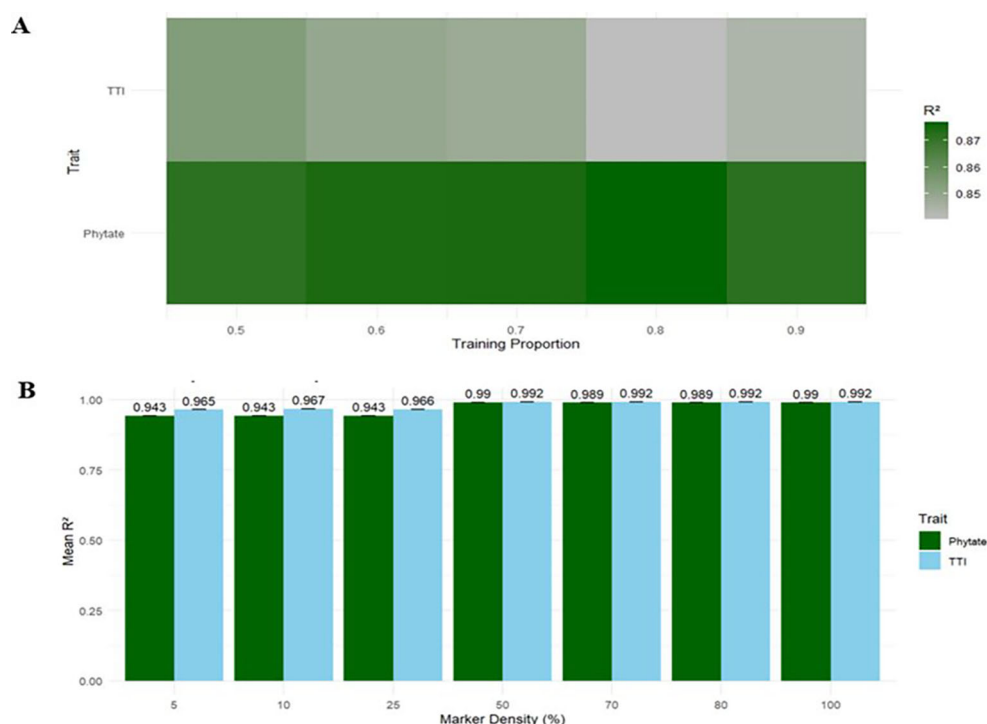


Figure 2. Regression curve for the predicted versus actual values for phytic acid and TTI using the 80:20 training-to-test set ratio.

Table 3. Training set performance for phytic acid and trypsin inhibitors using training set sizes between 0.5 and 0.9.

Proportion	Trait	R ²	RMSE	MAE	RPD
0.5	Phytic acid	0.87	409.34	305.73	2.80
0.5	TTI	0.85	105.89	80.08	2.63
0.6	Phytic acid	0.87	402.95	303.07	2.83
0.6	TTI	0.85	105.50	80.71	2.59
0.7	Phytic acid	0.87	415.98	310.31	2.83
0.7	TTI	0.85	108.15	83.34	2.57
0.8	Phytic acid	0.88	411.40	310.93	2.86
0.8	TTI	0.84	116.40	88.77	2.51
0.9	Phytic acid	0.87	433.89	324.71	2.80
0.9	TTI	0.84	115.91	89.11	2.54

R² – coefficient of determination; RMSE – root mean square error; MAE – mean absolute error; RPD – ratio of performance to deviation.

**Figure 3.** Heatmap of the coefficient of determination across training set proportions from 0.5 to 0.9 (A) and marker densities from 0.05 to 1.0 (B) for phytic acid and trypsin inhibitors using the Random Forest model.

Optimal combination for highest predictive accuracy was with training set of 0.7 or 0.8 with a marker density of 1.0 for phytic acid, and a training set of 0.5 with a marker density of 0.8 for TTI.

Analysis of variance (ANOVA) showed that training size and trait had a highly significant effect on model prediction accuracy ($p < 0.001$) as well as marker density ($p < 0.01$). The training set and marker density showed no significant interaction effects ($p > 0.01$) (Table 4).

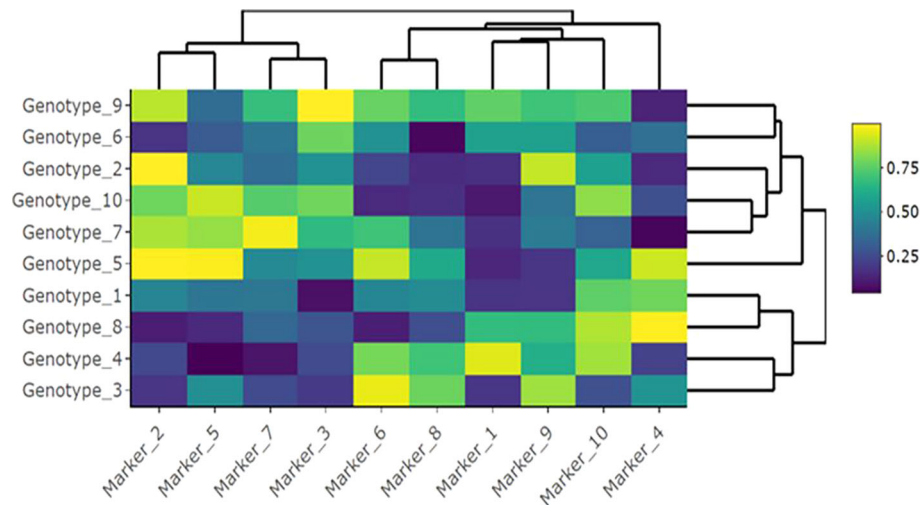
Correlation between significant markers and genotypes

SNP markers contributing to most of the variation across the genotypes were assessed. Markers were clustered into two main groups. The pairs clustered together were 2, 5, 7, and 3. The subsequent arms of the dendrogram clustered together with markers 6 and 8. Markers 1, 9, 10, and 4 clustered in the next arm. Higher correlations were observed between

Table 4. ANOVA for training set and marker density.

	df	Variance	F value	Pr(>F)	
Marker Density	6	0.0001	3.85	0.004	**
Proportion	4	0.0001	6.03	0.0008	***
Trait	1	0.0111	613.5	< 2e-16	***
Marker Density × Proportion	24	0.0000	0.37	0.992	ns
Residuals	34	0.0000			

Df – degree of freedom; Sum sq – sum of squares; Mean sq – mean squares; Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘.’ 1; ns – non significant. Lack of interaction observed between marker density and training set size suggests that increasing marker density has a similar effect across different training set sizes, and changing the training proportion does not significantly influence the effect of marker density.

**Figure 4. Correlation between the most significant markers and their corresponding genotypes.**

marker_3 and genotype 9, marker_2 and genotype 2, marker_2 and genotype 5, marker_5 and genotype 5, marker_7 and genotype 7, marker_6 and genotype 3, marker_1 and genotype 4, and marker_4 and genotype 8 (Figure 4).

Variable importance for prediction model performance

The percentage increase in the mean squared error (*%IncMSE*) and the increase in node purity (*IncNodePurity*) indicate that markers s_6_14309711, s_6_14300766, s_14255927, and s_19_40581884 are the most influential factors for model performance for phytic acid. For TTI, most influential predictors are s_4_47179496, s_7_5312274, s_11_100534382 and s_7_5620884 based on the *%IncMSE*, and s_11_10534382, s_4_47179496, s_17_35757744, and s_18_38890507 based on the *IncNodePurity*.

Variance components estimates for phytic acid and total trypsin inhibitors

The genotypic variance for phytic acid is 0.36 (SE = 0.15, Z-ratio = 2.36), while for TTI is 0.01 (SE = 0.01, Z-ratio = 1.29). For Residual variance, phytic acid has 0.95 (SE = 0.12, Z-ratio = 7.76) and TTI: 0.075486 (SE = 0.01, Z-ratio = 9.58) (Table 5).

Table 5. Variance components for phytic acid and trypsin inhibitors.

Trait	Genotypic variance	Residual variance	Z-ratio	Fixed effect estimate	AIC	BIC	H ²	h ²
Phytic acid	0.36 (SE = 0.15)	0.96 (SE = 0.1234)	2.36	1.76 (SE = 0.06, t = 31.49)	261.43	265.16	0.68	0.3
TTI	0.01 (SE = 0.01)	0.08 (SE = 0.01)	1.29	0.85 (SE = 0.02, t = 54.45)	305.3	309.03	0.84	0.1

#Akaike information criterion (AIC); Bayesian information criterion (BIC); Broad-sense heritability (H²); Narrow-sense heritability (h²); Standard error (SE).

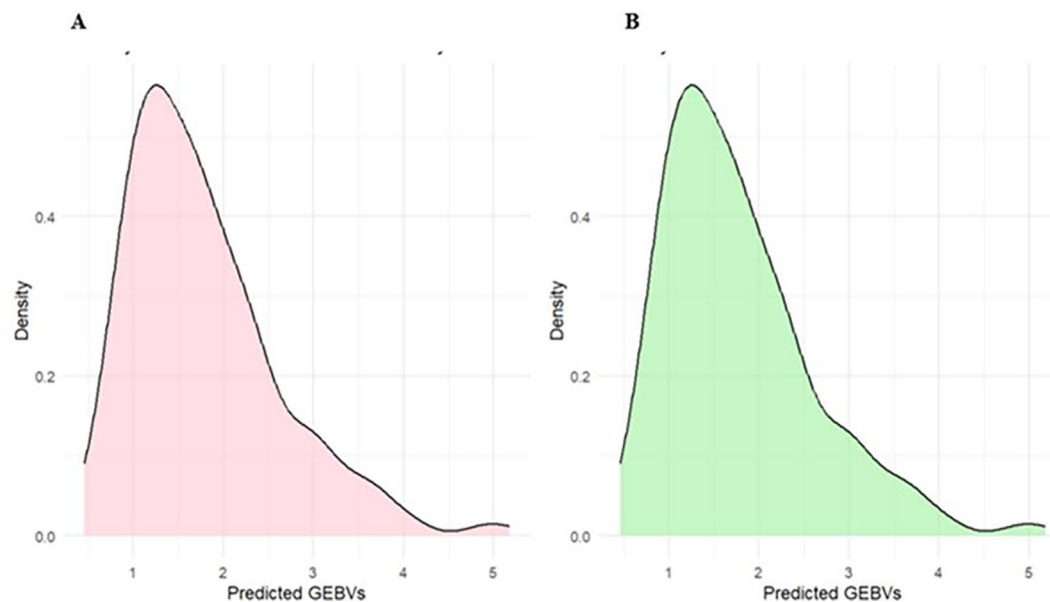


Figure 5. Density plots of predicted GEBVs for phytic acid (A) and trypsin inhibitors (B). Gaussian kernel density estimates for both phytic acid and TTI GEBVs are approximately at 1.5.

Genomic estimated breeding values (GEBVs) for phytic acid and TTI

The potential genetic contribution of an individual's genotype to phytic acid ranged from 0.45 to 5.18 (mean, 1.78, SE = 0.05), while for TTI, 0.35 to 1.25 (mean, 0.85, SE = 0.01) at 95% confidence level (Figure 5).

Discussion

This study assessed the statistical model, optimum training set size, and marker density for predicting phytic acid and trypsin inhibitors in soybeans. The average anti-nutritional factors (ANFs) observed were moderate, with phytic acid levels higher than the TTI in the genotypes. In contrast, it has been reported that TTI levels were higher than phytic acid in genetically modified (GM) cultivars.⁶⁴ The relatively lower content observed in our germplasm may be due to differences in the methodological approach and genotype improvement background, and over the years, efforts have been made to reduce anti-nutritional factors worldwide. Improvement strategies aimed at eliminating or reducing anti-nutritional factors may shape the genetic background of a genotype. Growing conditions affect ANFs content in soybeans.^{11,65,66} The range of anti-nutritional factor levels reported in this study could be a valuable opportunity for breeding soybean varieties with enhanced nutritional quality for feed and food.

In plant genetics research, emphasis has been placed on the importance of considering data distribution assumptions when analyzing phenotypic traits.⁵² Although the traits exhibited a normal distribution, the poor prediction performance observed using parametric models may suggest an underlying genetic complexity in the evaluated traits.^{67,68}

The weak positive linear relationship between these two traits suggests that changes in phytic acid levels are not highly predictive of changes in trypsin inhibitor levels. Thus, in breeding schemes, these two anti-nutritional factors could be considered for independent improvement. This may be beneficial for breeders to target trait-specific improvements.

The nonparametric model (Random Forest) outperformed the parametric models assessed in this study. Similarly, Random forest outperformed Partial Least Squares (PLS), Support Vector Regression (SVR),⁶⁹ and other machine learning models, including GBM and XgBoost⁷⁰ and deep learning models.²⁹ This supports the notion that Random Forest is a prominent model for dealing with nonlinear relationships and can handle highly complex traits such as ANFs in breeding schemes.

The slight differences in model performance based on the training set size and marker density suggest that the number of individuals and markers alone may not be sufficient to improve the accuracy. The training sets should be diverse for better model performance, as stated by a well-designed training set that accounts for its size, diversity, relatedness, and marker density.³⁹ This setup supports model improvement and avoids under- or over-representation.⁷¹ TSS has been reported as the main limiting factor for soybean accuracy. However, the rate of improvement decreases rapidly above a certain

number of individuals.^{72,73} Although marker density affects the prediction accuracy, in this study, no significant difference was observed above 50%. Likewise, the number of SNPs impacts prediction accuracy; however, no significant improvements beyond 1000 markers were observed.^{21,73} Therefore, trait-specific model optimization is crucial for traits with different genetic architectures.

The prediction model and GEBVs can be used to predict breeding values in a target set with only genotype data.²¹ The GEBVs suggested more potential for selecting genotypes with lower phytic acid levels compared to the narrower range for TTI, implying different breeding strategies for each trait. The Z-ratio showed substantial genetic variance for phytic acid, whereas it was moderate for TTI.

Selection based on GEBVs has a transformative impact on plant breeding by reducing the costs associated with phenotyping. Genomic prediction combined with machine learning will open new avenues for breeding schemes.

Conclusion

Random Forest (RF), a non-parametric machine learning model, can achieve superior predictive accuracy compared to traditional parametric linear models. Optimizing the training set size and marker density enhances genomic predictions up to a threshold, beyond which additional individuals and markers provide little to no benefit. The optimal training set size may vary depending on the trait, provided that the training set captures sufficient genetic variation to maximize the model performance. Additionally, increasing the number of markers continuously may lead to redundancy and diminishing returns on model performance.

Random Forest is a powerful tool for genotype selection based on genotypic data in breeding schemes that target the reduction of anti-nutritional factors in soybeans. Future research should focus on assessing the impact of linkage disequilibrium on prediction performance and incorporating environmental factors to further refine the model accuracy.

Geolocation

The experimental site, Kabanyolo, is geographically located within MUARIK, in the Central Region of Uganda at coordinates 0°28' N, 32°36' E.

Declaration of funding

The Regional Scholarship and Innovation Fund of the Partnership for Skills in Applied Sciences, Engineering, and Technology (RSIF-PASET) and the Carnegie Corporation of New York funded this work. The Genotyping cost was co-funded by the Bill and Melinda Gates Foundation [grant number OPP1093174] and the Integrated Genotyping Sequence Support (IGSS) project.

Author contributions

NP was responsible for collecting the soybean seeds, conducting the experiments and data analysis, and compilation of the manuscript; Dr. TO reviewed the first draft of the manuscript; Dr. JP assisted NP in the formal data analysis and reviewed the article; Dr. Isaac Onziga Dramadri was the project administrator; Dr. EN assisted in the data curation; Dr. MM assisted with the methodology and compilation of the graphs; Dr. MO reviewed the article; Dr. RE helped in funding management and reviewed the article; Prof. PT helped with the plant materials used in this study and revised the article.

AI Tools: Grammarly was used to check spelling and grammar.

Statement on agro-environmental ethics: This manuscript adheres to the agro-environmental ethics policy of the South African Journal of Plant and Soil.

Consent for publication

All authors read and approved the final version of the manuscript and agreed to its submission to the South African Journal of Plant and Soil (SAJPS).

ORCID IDs

Norberto Jose Palange: <https://orcid.org/0000-0003-3049-3224>

Julius Pyton Sserumaga: <https://orcid.org/0000-0001-5152-2835>

Tonny Obua: <https://orcid.org/0000-0003-1134-3255>

Dramadri Isaac Onziga: <https://orcid.org/0000-0002-9423-7126>

Data availability

All the necessary data are provided in the manuscript. Any other datasets will be available upon request.

Acknowledgements

The Partnership for Applied Skills in Sciences, Engineering, and Technology-Regional Scholarship and Innovation Fund (PASET-RSIF) and Carnegie Corporation of New York for funding this work. We appreciate the support provided by the Department of Crop Science and Horticulture for this work, Makerere Regional Center for Crops Improvement (MaRCCI) for providing the facilities for the Research and to Makerere University Centre for Soybean Improvement and Development (MAKCSID) staff for providing the soybean materials used in this research. Extended gratitude goes to the National Agricultural Research Organization (NARO) and technicians at the National Crops Resources Research Institute (NACRRI) for their support.

References

- Rosso ML, *et al.*: **Development of Breeder-Friendly KASP Markers for Low Concentration of Kunitz Trypsin Inhibitor in Soybean Seeds.** *Int. J. Mol. Sci.* 2021; **2675**: 1–16.
- Maria John KM, Khan F, Luthria DL, *et al.*: **Proteomic analysis of anti-nutritional factors (ANF's) in soybean seeds as affected by environmental and genetic factors.** *Food Chem.* 2017; **218**: 321–329.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Fan Y, Yang W, Yan Q, *et al.*: **Genome-Wide Identification and Expression Analysis.** *Genes (Basel).* 2020; **11**: 1–22.
[Publisher Full Text](#)
- Sultana MS, *et al.*: **Overexpression of soybean trypsin inhibitor genes decreases defoliation by corn earworm (*Helicoverpa zea*) in soybean (*Glycine max*) and *Arabidopsis thaliana*.** *Front. Plant Sci.* 2023; **14**: 1–16.
- Rada V, Lichovnikova M, Safarik I: **The effect of soybean meal replacement with raw full-fat soybean in diets for broiler chickens.** *J. Appl. Anim. Res.* 2017; **45**: 112–117.
[Publisher Full Text](#)
- DeMers LC, Raboy V, Li S, *et al.*: **Network Inference of Transcriptional Regulation in Germinating Low Phytic Acid Soybean Seeds.** *Frontiers Plant Sci.* 2021; **12**: 1–17.
[Publisher Full Text](#)
- Titarenko E, Rojo E, León J, *et al.*: **Jasmonic Acid-Dependent and Independent Signaling Pathways Control Wound-Induced Gene Activation in *Arabidopsis thaliana*.** *Plant Physiol.* 1996; **115**(81): 817–826.
[Publisher Full Text](#)
- Koiwa H, Bressan RA, Hasegawa PM: **Regulation of protease inhibitors and plant defense.** *Trends Plant Sci.* 1997; **2**: 379–384.
[Publisher Full Text](#)
- Martins V, Ferrari F, White PJ: **Plant Physiology and Biochemistry Phytic acid accumulation in plants: Biosynthesis pathway regulation and role in human diet.** *Plant Physiol. Biochem.* 2021; **164**: 132–146.
[Publisher Full Text](#)
- Kumar V, *et al.*: **Phytic acid in Indian soybean: Genotypic variability and influence of growing location.** *J. Sci. Food Agric.* 2005; **85**: 1523–1526.
[Publisher Full Text](#)
- Shiguro TI, No TO, Ada TW, *et al.*: **Changes in Soybean Phytate Content as a Result of Field Growing Conditions and Influence on Tofu Texture.** *Biosci. Biotechnol. Biochem.* 2006; **70**: 874–880.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Raboy V: **Approaches and challenges to engineering seed phytate and total phosphorus.** *Plant Sci.* 2009; **177**: 281–296.
[Publisher Full Text](#)
- Lush JL, Hazel LN: **The efficiency of three methods of selection.** *J. Hered.* 1942; **33**: 393–399.
- Vieira RA, Rocha R, Scapim CA, *et al.*: **Selection index based on the relative importance of traits and possibilities in breeding popcorn.** *Genet. Mol. Res.* 2016; **15**: 1–11.
- Henderson CR: **Best linear unbiased estimation and prediction under a selection model published by: international biometric society stable.** *Biometrics.* 1975; **31**: 423–447.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wakchaure R, Ganguly S: **Marker Assisted Selection (MAS) in Animal Breeding: A Review.** *J. Drug Metab. Toxicol.* 2015; **06**: 6–9.
[Publisher Full Text](#)
- Riedelsheimer C, *et al.*: **Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize.** *PNAS.* 2012; **109**: 8872–8877.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cobb JN, Biswas PS, Platten JD: **Back to the future: revisiting MAS as a tool for modern plant breeding.** *Theor. Appl. Genet.* 2019; **132**: 647–667.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang CL, *et al.*: **Comparison of five methods for genomic breeding value estimation for the common dataset of the 15th QTL-MAS Workshop.** *BMC Proc.* 2012; **6**: 2–6.
- Hickey JM, *et al.*: **Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation.** *Crop Sci.* 2015; **54**: 1476–1488.
- Kaler AS, Purcell LC, Beissinger T, *et al.*: **Genomic prediction models for traits differing in heritability for soybean, rice, and maize.** *BMC Plant Biol.* 2022; **22**: 11–87.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bhat JA, *et al.*: **Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding.** *Front. Genet.* 2016; **7**: 1–11.
- Davey JW, *et al.*: **Genome-wide genetic marker discovery and genotyping using next-generation sequencing.** *Nat. Publ. Gr.* 2011; **12**: 499–510.
- Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics.* 2001; **157**: 1819–1829.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sun B, *et al.*: **Genetic variation and marker–trait association affect the genomic selection prediction accuracy of soybean protein and oil content.** *Front. Plant Sci.* 2022; **13**: 1–10.
- Tamba CL, Ni YL, Zhang YM: **Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies.** *PLoS Comput. Biol.* 2017; **13**: 1–20.
[Publisher Full Text](#)
- Kayanan M, Wijekoon P: **Performance of LASSO and Elastic net estimators in Misspecified Linear Regression Model.** *Ceylon J. Sci.* 2019; **48**: 293.
- Liu L, Gao J, Beasley G, *et al.*: **LASSO and Elastic Net Tend to Over-Select Features.** *Mathematics.* 2023; **11**: 1–16.
[Publisher Full Text](#)

29. Gill M, *et al.*: Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. *BMC Plant Biol.* 2022; **22**: 1–8.
30. Yoosefzadeh-Najafabadi M, *et al.*: Machine-Learning-Based Genome-Wide Association Studies for Uncovering QTL Underlying Soybean Yield and Its Components. *Int. J. Mol. Sci.* 2022; **23**.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Norman A, Taylor J, Edwards J, *et al.*: Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *Genes, Genome Genet.* 2018; **8**: 2889–2899.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Berro I, Lado B, Nalin RS, *et al.*: Training Population Optimization for Genomic Selection. *Plant Genome.* 2019; **12**: 1–14.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Wu PY, Ou JH, Liao CT: Sample size determination for training set optimization in genomic prediction. *Theor. Appl. Genet.* 2023; **136**: 57.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Nsibi M, *et al.*: Adoption and optimization of genomic selection to sustain breeding for apricot fruit quality. *G3 Genes, Genomes, Genet.* 2020; **10**: 4513–4529.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Montesinos-López OA, *et al.*: Efficacy of plant breeding using genomic information. *Plant Genome.* 2023; **16**: 1–14.
[Publisher Full Text](#)
36. Zhu S, *et al.*: Evaluation of Bayesian alphabet and GBLUP based on different marker density for genomic prediction in Alpine Merino sheep. *G3 Genes, Genomes, Genet.* 2021; **11**.
37. Isidro J, *et al.*: Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* 2015; **128**: 145–158.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Rolling WR, Lake R, Dorrance AE, *et al.*: The Effects of Genetic Distance and Genetic Diversity on Genomic Prediction Accuracy for Soybean Quantitative Disease Resistance to *Phytophthora sojae*. *PhytoFrontiers™*. 2024; **4**: 272–276.
[Publisher Full Text](#)
39. Sallam AH, Endelman JB, Jannink J-L, *et al.*: Assessing Genomic Selection Prediction Accuracy in a Dynamic Barley Breeding Population. *Plant Genome.* 2015; **8**: eplantgenome2014.05.0020.
[PubMed Abstract](#) | [Publisher Full Text](#)
40. Obua T, *et al.*: Yield stability of tropical soybean genotypes in selected agro-ecologies in Uganda. *South African J. Plant Soil.* 2020; **37**: 168–173.
[Publisher Full Text](#)
41. R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2024. (2024).
[Reference Source](#)
42. Kilian A, *et al.*: Diversity arrays technology: A generic genome profiling technology on open platforms. *Methods Mol. Biol.* 2012; **888**: 67–89.
[PubMed Abstract](#) | [Publisher Full Text](#)
43. Egea LA, Mérida-García R, Kilian A, *et al.*: Assessment of genetic diversity and structure of large garlic (*Allium sativum*) germplasm bank, by diversity arrays technology 'genotyping-by-sequencing' platform (DARTseq). *Front. Genet.* 2017; **8**: 1–9.
44. Baloch FS, *et al.*: A whole genome DARTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. *PLoS One.* 2017; **12**: 1–18.
[Publisher Full Text](#)
45. Israel DW: Genetic Variability for Phytic Acid Phosphorus and Inorganic Phosphorus in Seeds of Soybeans in Maturity Groups V, VI, and VII. 2013; **46**: 67–71.
46. Anozie AN, Salami OA, Babatunde DE, *et al.*: Comparative evaluation of processes for production of soybean meal for poultry feed in Nigeria Evaluación comparativa de procesos para la producción de harina. *Anim. Sci.* 2018; **52**: 193–202.
47. Palange NJ, *et al.*: Genetic variability of anti - nutritional factors among soybean (*Glycine max.*) germplasm. *Discov. Agric.* 2025; **3**: 1–14.
48. Palange NJ, *et al.*: Artificial intelligence - driven near - infrared spectrophotometry model for rapid quantification of anti - nutritional factors in soybean (*Glycine max.*). *Discov. Appl. Sci.* 2025; **7**: 1–13.
49. Wickham H, François R, Henry L, *et al.*: *A Grammar of Data Manipulation. R package version 1.1.3. dplyr.* 2023. 2023.
[Reference Source](#)
50. Wickham H: *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2023.
51. John F, Weisberg SA: *{R} Companion to Applied Regression*. Thousand Oaks CA: Sage; Third ed. 2019; vol. **2019**.
52. Howard R, Carriquiry AL, Beavis WD: Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 Genes, Genomes, Genet.* 2014; **4**: 1027–1046.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. Habier D, Fernando RL, Kizilkaya K, *et al.*: Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics.* 2011; **12**.
54. Endelman J: Ridge Regression and Other Kernels for Genomic Selection. *Plant Genome.* 2011; **4**: 250–255.
55. Pérez P, De Los Campos G: Genome-wide regression and prediction with the BGLR statistical package. *Genetics.* 2014; **198**: 483–495.
[PubMed Abstract](#) | [Publisher Full Text](#)
56. Friedman J, Hastie T, Tibshirani R: Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 2010; **33**: 1–22.
[PubMed Abstract](#) | [Publisher Full Text](#)
57. Liaw A, Wiener M: The R Journal: Classification and regression by randomForest. *R J.* 2002; **2**: 18–22.
58. Wang X, *et al.*: Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese simmental beef cattle. *PLoS One.* 2019; **14**: 1–14.
[Publisher Full Text](#)
59. Desta ZA, Ortiz R: Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* 2014; **19**: 592–601.
[PubMed Abstract](#) | [Publisher Full Text](#)
60. Hastie T, Tibshirani R, Friedman J: *Statistics The Elements of Statistical Learning. Math. Intell.* 2009; **27**: 83–85.
61. Wickham H: Reshaping data with the reshape package. *J. Stat. Softw.* 2007; **21**: 1–20.
[Publisher Full Text](#)
62. Bohlouli M, Alijani S, Javaremi AN, *et al.*: Genomic prediction by considering genotype × environment interaction using different genomic architectures. *Ann. Anim. Sci.* 2017; **17**: 683–701.
[Publisher Full Text](#)
63. Covarrubias-Pazarán G: Genome-Assisted prediction of quantitative traits using the r package sommer. *PLoS One.* 2016; **11**: 1–15.
[Publisher Full Text](#)
64. Galão OF, *et al.*: Acta Scientiarum Kunitz trypsin Inhibitor and phytic acid levels in conventional and genetically modified soybean seeds from Londrina and Ponta Grossa, South Brazil. 2014; 727–731.
[Publisher Full Text](#)
65. Israel DW, Kwanyuen P, Burton JW, *et al.*: Response of low seed phytic acid soybeans to increases in external phosphorus supply. *Crop Sci.* 2007; **47**: 2036–2046.
[Publisher Full Text](#)
66. Vollmann J, Grausgruber H, Wagentrists H, *et al.*: Trypsin inhibitor activity of soybean as affected by genotype and fertilisation. *J. Sci. Food Agric.* 2003; **83**: 1581–1586.
[Publisher Full Text](#)
67. Chun S, *et al.*: Non-parametric Polygenic Risk Prediction via Partitioned GWAS Summary Statistics. *Am. J. Hum. Genet.* 2020; **107**: 46–59.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
68. Zeng P, Zhou X: models. *Nat. Commun.* 2017; **8**: 1–11.
69. Robinson RLM, Palczewska A, Palczewski J, *et al.*: Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* 2017; **57**.
70. Li B, *et al.*: Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* 2018; **9**: 1–20.
71. Alemu A, *et al.*: Genomic selection in plant breeding: Key factors shaping two decades of progress. *Mol. Plant.* 2024; **17**: 552–578.
[PubMed Abstract](#) | [Publisher Full Text](#)
72. Xavier A, Muir WM, Rainey KM: Assessing predictive properties of genome-wide selection in soybeans. *G3 Genes, Genomes, Genet.* 2016; **6**: 2611–2616.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
73. Beche E, *et al.*: Genomic prediction using training population design in interspecific soybean populations. *Mol. Breed.* 2021; **41**: 15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)