

RESEARCH

Open Access



Genomic prediction accuracy for low hydrogen cyanide selection in fresh cassava roots. Comparative model analysis

Michael Kanaabi^{1,2*} , Settumba B. Mukasa¹ , Ephraim Nuwamanya¹ , Ismael Siraj Kayondo³ , Henry Wagaba⁴ , Paula Iragaba⁵ , Julius K. Baguma⁵ , Nicholas Muhumuza⁵, Ann Ritah Nanyonjo^{1,5} , Enoch Wembabazi⁵ , Mary Buttibwa⁵, Williams Esuma⁵ and Robert Kawuki⁶

Abstract

Cassava (*Manihot esculenta* Crunz) is a staple food crop for millions of people in Africa. The crop's tolerance to drought, ability to survive in marginal soils and amenability to a variety of uses endears it to farmers. However, utilization of the crop's full potential for food security is limited by the presence of cyanogenic glucosides (HCN) in the roots. These not only make cassava bitter to the taste but can also be toxic to lethal levels. Breeding for low HCN cassava is necessary to mitigate against high dietary HCN consumption. However, efforts at variety development have been slow, in part due to limited use of molecular tools. This study assessed genomic prediction accuracy for fresh cassava root HCN content using the GBLUP model and compared its performance to 5 Bayesian models (Bayes A, Bayes B, Bayes C, Bayesian Ridge Regression (BRR), Bayesian Lasso (BL) and the Reproducing Kernel Hilbert Spaces (RKHS) model. We used a genomic selection (GS) cycle two population consisting of 434 clones which were genotyped with 24,040 SNP markers using the DArTseq platform. Data were filtered to remove markers with minor allele frequency (MAF) less than 0.05, call rate less than 80%, clones with more than 20% missing calls and clones with no matching phenotypic data. Phenotypic data on fresh cassava root HCN content was collected on roots at 12 months after planting. The accuracy of GBLUP model was low, ranging from -0.41 to 0.68 and averaging 0.22 across folds. On the other hand, the accuracy of RKHS and Bayesian models was moderate, with BL and RKHS being most accurate ($r=0.52$), followed by Bayes A ($r=0.49$) and Bayes B and Bayes C ($r=0.48$). The BRR model was least accurate ($r=0.18$). Collectively, these results provide a foundation for implementing GS to accelerate efforts to deliver desired genetic gains for low HCN cassava in farmers' fields.

Keywords Food safety, Genetic gain, Genomic selection, Hydrogen cyanide, Marker assisted selection

*Correspondence:

Michael Kanaabi
kanaabimichael@gmail.com

¹College of Agricultural and Environmental Sciences, Makerere University,
P.O. Box 7062, Kampala, Uganda

²National Coffee Research Institute (NaCORI), P.O. Box 185, Mukono,
Uganda

³International Institute for Tropical Agriculture (IITA), Ibadan
200113, Nigeria

⁴National Agricultural Research Laboratories (NaRL), P.O. Box 7065,
Kampala, Uganda

⁵National Crops Resources Research Institute (NaCRRI), P.O. Box 7084,
Kampala, Uganda

⁶World Coffee Research (WCR), 1094 SW Barnes Road #334, Portland,
OR 97225, USA



Background

Cassava is a staple food crop for millions of people in Africa, Asia and South America. In addition to having multiple uses for food, feed and industry, the crop is hardy and survives well in marginal soils and under drought conditions [1, 2], making it a major food security crop for resource limited farmers in the face of changing climatic conditions. However, the usefulness of the crop is limited by the presence of the cyanogenic glucosides linamarin and lotaustralin in the leaves and roots. In neutral or alkaline media and in presence linamarase enzyme, linamarin readily hydrolyses into glucose and acetone cyanohydrin. The acetone cyanohydrin decomposes to liberate cyanide and a hydrogen ion (HCN) [3, 4]. In the cassava plant, linamarin and lotaustralin are stored in the cell vacuoles while linamarase is stored in the cell wall. This spatial separation ensures that cyanide release does not occur in intact tissues. However, when root tissues are damaged say by peeling, grating or chewing, cell structures are broken down, allowing interaction of linamarin with linamarase, leading to liberation of volatile hydrogen cyanide (HCN) [3]. This HCN makes the cassava roots bitter [5]. Chronic dietary exposure to HCN is associated with the neurological disorders; konzo [6, 7] and tropical ataxic neuropathy [8, 9] while ingestion of high HCN cassava can be acutely toxic [10, 11]. Whereas communities traditionally cultivating high HCN cassava varieties know processing techniques for HCN reduction, sometimes the processing is not effective [12] and when these roots are sold in distant markets, buyers may not always be aware of the potentially lethal levels of HCN in the roots [10, 11]. The most sustainable way to guard the public from harmful effects of dietary HCN consumption is by breeding for low HCN cassava varieties that meet other user preferences. However, breeding efforts have been slow due to challenges associated with HCN phenotyping as current methods [13–16] are slow, laborious and/or limited in accuracy.

A significant challenge in cassava breeding is the crop's long cycle time, which can take up to 12 months to reach physiological maturity. For traits such as fresh cassava root HCN content, which are evaluated at physiological maturity, it can take a full year to advance from one breeding stage to the next. Moreover, the ability to select individuals at early breeding stages is crucial for maximizing genetic gains and accelerating the development of new varieties, making it a key objective for cassava breeding programs [17]. Genomic selection (GS) leverages genome-wide markers to predict an individual's genetic merit, enabling selection decisions to be made earlier [18, 19]. The availability of genome-wide markers makes GS applicable for traits controlled by a few major genes like the oligogenic HCN [20] as well as those influenced by many small-effect genes. Compared to traditional

phenotypic selection, GS allows for selection before phenotyping, reducing costs and, in some cases, proving more accurate than marker-assisted selection [21, 22]. Whereas the potential of deploying marker assisted selection for low HCN cassava breeding has been variously demonstrated [23–26], only one published study [25] evaluated the possibility of genome-wide markers assisted selection for HCN to facilitate germplasm exchange between breeding programs.

Historically, the Uganda cassava breeding program has employed genomic selection for the cyclic improvement of traits such as resistance cassava brown streak disease [19, 27, 28], dry matter content, and total carotenoids [18]. Prediction accuracies, achieved using the genomic best linear unbiased predictor (GBLUP) and Bayesian models, have ranged from -0.16 to 0.68 [18, 19, 27, 28]. The GBLUP model utilizes a genomic relationship matrix (GRM) built from SNP marker data to capture the genetic relationships between individuals. It applies a linear mixed model where the genetic effects, represented by the GRM, are treated as random effects [29]. The inverse of the genomic relationship matrix was calculated using the *G.inverse* function and used to account for population structure and relatedness. The model then predicts the genetic value of individuals by aggregating the effects of individual markers across the genome. The GRM enables more precise estimation of genetic relationships, resulting in improved predictions [30]. Additionally, it accounts for linkage decay, which refers to the non-random association of alleles at different loci, thereby enhancing prediction accuracy. Bayes A model assumes that marker effects follow a *t*-a prior distribution and shrinks the estimates of marker effects towards zero, thereby reducing the risk of overfitting. Marker effects are assumed to be independent of each other, and a few markers are assumed to have large effects while many others have small effects [31]. Similar to Bayes A, Bayes B assumes that marker effects are independent and that most markers have no influence on the trait. The markers that influence the trait are assumed to follow a scaled-*t* prior distribution [32]. In contrast, Bayes C operates under the assumption that only a small proportion of SNPs have moderate to large effects, while the majority have no effect. Additionally, Bayes C integrates prior knowledge about the distribution of SNP effects [32]. In contrast, Bayesian Ridge Regression (BRR) assigns a Gaussian prior with a shared variance to each marker effect, applying uniform shrinkage across all marker effects [31]. On the other hand, the Bayesian Lasso (BL) model assumes that SNP effects are independent and uses a double-exponential prior distribution for marker effects, which heavily shrinks markers with little or no impact on the trait [33, 34]. The primary differences between the models lie in their assumptions regarding marker effects.

In Reproducing Kernel Hilbert Spaces (RKHS), a Gaussian kernel serves as a basis function to estimate conditional expectations. The goal is to reduce numerous genetic markers into a positive semi-definite matrix of size $n \times n$, where n represents the number of phenotypes (HCN scores), by constructing genetic kinship based on spatial distance within a specific metric space. For this study, the Gaussian kernel was utilized.

$$K_{ij} = \exp(-(d_{ij}\theta))$$

Where K_{ij} is the computed relationship between two clones, d_{ij} is the Eucladian genetic distance based on marker dosages and θ is a tuning parameter that determines the rate of decay of correlation among clones.

The performance of the GBLUP model has been comparable to or better than that of the Bayesian models. However, there is no universally optimal genomic prediction model, as a set of assumptions that work well in one scenario may not perform effectively in another [35]. The accuracy of genomic predictions can be affected by the genetic architecture of the trait, the trait heritability, marker density, size of the training set, population structure, genetic variance and choice of genomic prediction model among others [27, 28, 36–38]. Therefore, it is essential to evaluate multiple models to determine which one is best suited for the HCN dataset.

Whereas this genomic selection cyclic population has been studied for cassava mosaic and cassava brown streak disease no similar work has been done for root quality traits, despite root quality being a major consideration in selection of parental genotypes in the genomic selection cycle zero (C0) population. The phenotyping of cassava root HCN content is complex given that it is costly, slow and highly labor-intensive. For example, using the picrate paper method [39], one has to prepare picrate paper strips which require at least 24 h of air drying after dipping them in picrate solution (solution of picric acid and sodium carbonate), dig roots out of the ground, wash to remove debris, peel to remove the cortex, weigh off a sample for the analysis, expose the sample to the picrate paper strip, incubate in the dark and wait for 12 to 16 h before matching the color change on the paper strip to a color chart. This difficulty in HCN phenotyping coupled with the fact that the trait is only evaluated after physiological maturity, makes optimizing genomic predictions for HCN a high priority as it would accelerate genetic gains for low HCN cassava breeding while significantly reducing the cost of selecting low HCN cassava accessions. Besides, information on genomic predictions for HCN in fresh cassava roots is limited. Accordingly, this study aimed at assessing the possibility of using GS as a breeding

tool for selection of accessions with low HCN, an intricate and yet desirable quality trait. This possibility was done through undertaking two empirical studies to; 1) assess genomic prediction accuracy for fresh cassava root HCN content using the GBLUP model 2) compare performance of the GBLUP model to 5 Bayesian models (Bayes A, Bayes B, Bayes C, Bayesian Ridge Regression (BRR), Bayesian Lasso (BL) and the Reproducing Kernel Hilbert Spaces (RKHS) model. It is justified to evaluate several models since no single model is best fit for all data sets [31, 40]. Moreover, different genomic prediction models impose different computational demands on the breeding program. Thus, for resource limited publicly funded breeding programs, identifying models that are computationally efficient improves their efficiency given that GS often involves thousands of individuals and thousands of markers. Moreover, increasing model complexity can incur huge computational costs without yielding proportional gains in predictive accuracy [41].

Materials and methods

The study population

The study material consisted of a genomic selection cycle two (C2) clonal population [27]. This population originated from a cycle zero (C0) genomic selection population, which included 348 clones derived from forty-nine progenitors obtained from the National Crops Resources Research Institute (NaCRRI), the International Institute of Tropical Agriculture (IITA), and the International Centre for Tropical Agriculture (CIAT) in 2013. Progenitors from CIAT were selected based on their root quality performance, while those from NaCRRI and IITA were chosen for their fresh root yield, resistance to cassava brown streak disease (CBSD), and root quality traits. Elite progenitors from this group were identified using their genomic estimated breeding values (GEBVs) and used to develop a cycle one (C1) population comprising 638 clones. Subsequently, elite progenitors from C1 were crossbred to produce 6,570 half-sib and full-sib seedlings, forming the genomic selection cycle two (C2) population. These seedlings were established in a field at NaCRRI in 2018 to multiply planting material.

Due to devastation by cassava brown streak disease, only 434 seedlings survived with sufficient planting material to establish multi-location trials (Fig. 1). These were cloned and established in field trials in Namulonge (2019, 2020), Serere (2019, 2020, 2021) and Ngetta (2021) in an augmented design [18] consisting of 29 blocks with 2 checks per block. Each block had 15 plots (accessions), with each plot consisting of a single row of 10 plants. The plants were established at a spacing of one meter between plants in a row and one meter between adjacent rows. Routine

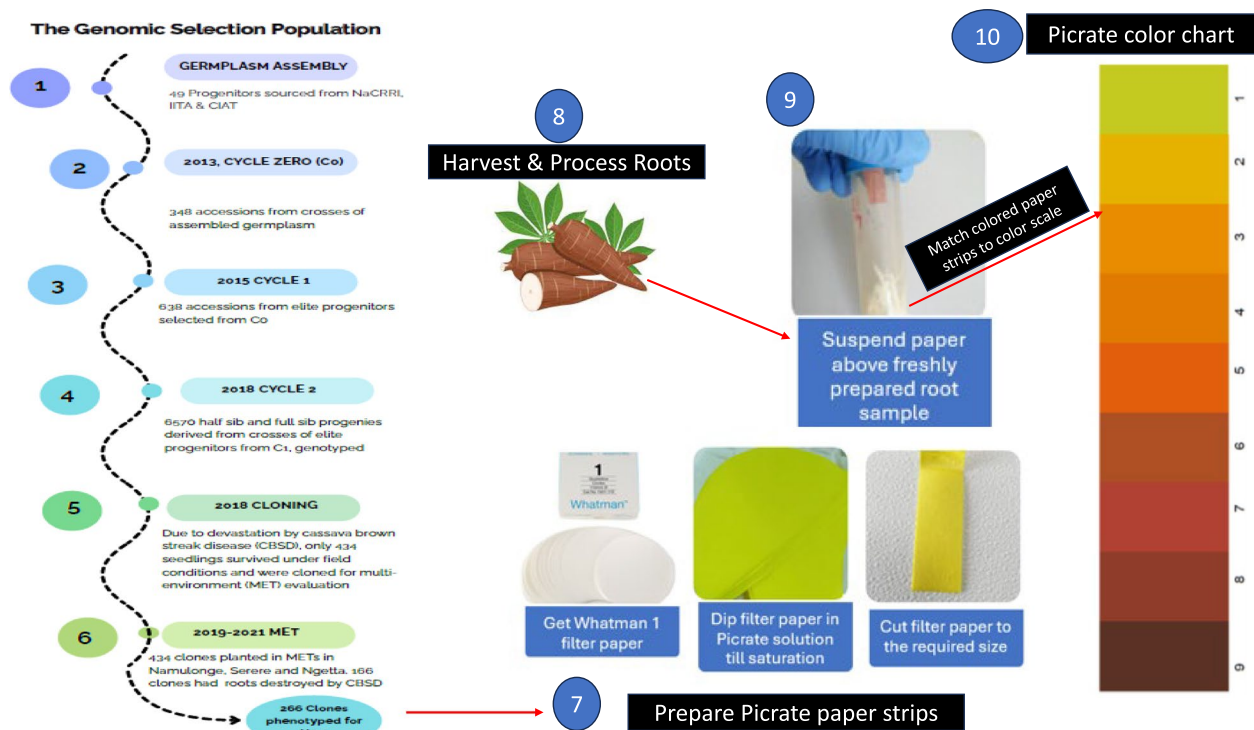


Fig. 1 Workflow for generating the cassava genomic selection population and phenotyping cassava roots for HCN using the Picrate paper color scale protocol

weeding was done using a hand-hoe to maintain the fields weed free. In the end, phenotypic data was collected from only 266 clones since we could not obtain sufficient roots from 168 clones owing to damage by cassava brown streak disease necrosis, which rendered them unsuitable for evaluation for HCN content [42].

Sampling for genotyping, DArTseq genotyping and SNP quality control

Leaf sampling for genotyping was done three months after planting in the field trial established in at NaCCRI in Namulonge. Two young fully opened leaves were identified and four-leaf punches collected from these using a 5 mm cork borer. Fresh leaves were placed in a 96-well genotyping plate, dried using silica gel in a desiccator. DNA extraction, genotyping, and SNP calling were performed using the DArTseq platform (<https://www.universityarrays.com/services/dartseq/>). The resultant sequences were aligned to cassava reference genome version 6.1. A total of 240,40 markers were called. The SNPs were then filtered to remove those with minor allele frequency less than 0.05, call rate less than 80% and more than 20% missing calls using the *qc.filtering* function in *ASRgenomics* R package [43]. After filtering, 4594 SNPs with MAF less than 0.05 and 10 clones with more than 20% missing calls were removed, leaving 19,446 markers and 255 clones that were used for downstream analysis.

Phenotypic data collection

At maturity, 12 months after planting, the roots of the five middle plants were harvested and combined. Three healthy (non-necrotic), uniformly sized roots were selected and transported to the laboratory. The roots were rinsed under running water to remove soil and debris, then dried with a kitchen towel. The extreme ends (distal and proximal ends) of each root were cut off and discarded, leaving a 5 cm middle section. This middle portion was peeled and grated using a kitchen grater. The grated material from each root was wrapped together, and a sample was immediately weighed into a falcon tube containing 5 mL of pH 8 phosphate buffer, with duplicates prepared for each sample, making 6 observations per plot. A fresh picrate paper strip (1 cm wide, 3 cm long), prepared by dipping Whatman 1 filter paper in a yellow solution of picric acid (0.5% w/v) and sodium carbonate (2.5% w/v) and allowing the paper to dry for 24 h at room temperature was suspended above the falcon tube, and the tube was tightly sealed. The tubes were incubated in the dark for 16 h at room temperature and scored on a scale of 1 to 9, where 1 and 9 represent extremes of low and high HCN respectively (Fig. 1) [39]. Because HCN phenotyping is labor intensive and time consuming, only 20 randomly selected plots were processed per day. The 6 observations per plot were averaged to get a mean plot HCN score which was used for downstream analysis.

Data analysis

Phenotypic data analysis

The data collected over various locations and years was combined into a single file. Histograms were created using the *ggplot2* package in R to visualize the data distribution. A mixed linear model was then fitted using the *ASReml-R* package in R, which estimated variance components via the restricted maximum likelihood method [44]. These variance components were subsequently used to calculate broad-sense heritability (H^2). The linear mixed model fitted was as below:

$$y = X\beta + Z_{clone}c + Z_{block}b_{location} + Z_{clone}l_{location} + \varepsilon$$

Where y represents the HCN phenotype, β is the fixed effect for the population mean with X the corresponding incidence matrix. The incidence matrix Z_{clone} and vector c represent fixed effects of clones where $C \sim N(0, I\sigma^2_c)$ with I being the identity matrix. The block nested within location effects were modeled as a random term with incidence matrix Z_{block} and effects vector $b \sim N(0, I\sigma^2_b)$. Vector l and incidence matrix ($Z_{clone}l_{location}$) represent the random effect of clone by location interaction and ε term represents the residual, distributed as $\varepsilon \sim N(0, \sigma^2_\varepsilon)$.

Broad-sense heritability (H^2) estimates were computed using the formula:

$$H^2 = \frac{\sigma^2_c}{\sigma^2_c + \sigma^2_{cl} + \sigma^2_e}$$

Where; σ^2_c represents the variance attributed to the clone, σ^2_{cl} denotes the variance due to the clone by location interaction, σ^2_e is the residual variance. Narrow-sense heritability (h^2) was calculated from the genomic relationship matrix via the formula:

$$h^2 = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_e}$$

Where; σ^2_a represents the additive genetic variance, σ^2_e denotes the non-additive genetic variance. The model was used to obtain Best Unbiased Linear Estimates (BLUEs) of each clone.

Assessment of population structure and kinship

After filtering genotype data and removing clones with more than 20% missing calls, 19,446 SNPs and 255 clones were retained and used for downstream analyses. The SNPs were formatted as a dosage matrix with genotypes coded as 0, 1 or 2 which represented alternative allele homozygotes, heterozygotes and reference allele

homozygotes respectively using the *snp.recode* function in *ASReml-R* package. A kinship matrix was generated from a genomic relationship matrix (G) as described by Van Raden [29]. Let M be a marker matrix specifying marker alleles that each individual inherited. The dimensions of M are given by the number of individuals (n) by the number of loci (m). The genetic relationships between markers were evaluated using an $n \times n$ matrix MM' . The genomic relationship matrix G was then calculated from the formula,

$$G = \frac{ZZ' / 2 \sum P_i (1 - P_i)}$$

Where Z_i is the difference between marker matrix M and the allele frequency (P). The allele frequency was expressed as the difference from 0.5 multiplied by 2 such that column i of P is $2(P_i - 0.5)$.

The kinship matrix was analyzed using the *kinship.diagnostics* function, and a plot of the off-diagonal values was generated using the *plot.offdiag* function. The diagonal elements of MM' represent the inbreeding coefficient for each individual, while the off-diagonal elements measure the number of alleles shared among relatives (genetic relatedness among pairs of individuals). A heatmap of the genomic relationship matrix was plotted to visualize the population structure using the *kinship.heatmap* function. Additionally, principal component analysis (PCA) was conducted using the *kinship.pca* function from the *ASReml-R* package, and the first two principal components were plotted to further illustrate the population structure.

Genomic prediction models

The BLUEs obtained from phenotype data were used as a response variable to perform genomic prediction. The GBLUP model was implemented using the licensed software *ASReml-R* and the estimated breeding values were extracted using the *predict* function [45]. We fitted the model;

$$Y = 1\beta + X_g g + \varepsilon \quad \text{with} \quad g \sim N(0, K\sigma_g^2) \quad \text{and} \quad \varepsilon \sim (0, I\sigma_\varepsilon^2)$$

Where; Y is a vector of BLUEs, β is the overall population mean, X is a design matrix linking observations to genomic values, g is a vector of genomic estimated breeding values for each clone and ε is a vector of residuals. We made the assumption that g had a known covariance structure, defined by the realized genomic relationship matrix K and identity matrix I .

Predictions using RKHS and Bayesian models were conducted with the BGLR package [46] in R. For the Bayesian methods, the same model as GBLUP was applied, assuming that the vector of estimated breeding

values for each clone (g) has an established covariance structure defined by the realized genomic relationship matrix and the identity matrix. The Bayesian methods were implemented using Markov Chain Monte Carlo (MCMC) sampling, incorporating the respective prior distributions for marker effects. For model tuning, the *burnIn* hyperparameter, which specifies the number of iterations at the beginning of the MCM chain that were ignored to remove bias from initial values was set to 700, the *nIter* hyperparameter which controls the total number of iterations of the MCMC sampler was set to 6000, the degrees of freedom (*df0*) were set to 5 while the scale parameter (*S0*) was set to 500.

Cross validation

A k-fold cross-validation approach was employed to assess the model's performance. The population was randomly divided into five subsets, and this process was repeated five times. During each cross-validation iteration, four of the five subsets were used as the training set, while the remaining one-fifth served as the validation set for model testing. Narrow-sense heritability was estimated separately within each validation set by portioning the phenotypic variance into genetic and error term components. The variance components were sampled from scaled inverse-chi square priors via MCMC. For each fold, posterior samples obtained after burn-in were averaged to yield the mean narrow-sense heritability for that fold. Genomic prediction accuracy was recorded for each iteration, and the process was repeated until every subset had been used as the test set, resulting into a total of 25 correlations. The model's predictive ability (r_a) was evaluated based on the correlation between the BLUES and the predicted values. Accuracy, on the other hand, was calculated by dividing the predictive ability (r_a) by

the square root of the narrow-sense heritability of the validation set [47, 48].

Results

Phenotypic variability

The HCN phenotypes across locations and years were scored on a 1 to 9 scale. In Namulonge, the HCN phenotypic scores ranged from 1 to 9, averaging 5.3. In Ngetta, the scores ranged from 4 to 9, with an average of 6.4 while in Serere, HCN scores ranged from 1 to 9, averaging 6.6. Whereas HCN scores in Namulonge approximated a normal distribution, scores in Serere and Ngetta were skewed towards the high values (Fig. 2). The overall mean score across locations was 6.1. The broad-sense heritability (H^2) across locations was 0.42. Overall, genomic selection is appropriate as the trait displays phenotypic variability.

Population structure and kinship

The SNP markers were distributed across all 18 chromosomes of the cassava genome (Fig. 3A). Population structure was investigated using a genomic relationship matrix and principal component analysis (PCA). A plot of genomic relationship matrix revealed kinship (relatedness) among some of the clones. Whereas the diagonal of the plot represents the kinship coefficient of an individual with itself, the off diagonal represents the kinship coefficient between different individuals in the population. The strength of the genetic relationship between individuals is signified by the color intensity from blue (no kinship) to yellow (full siblings) or parent-off spring relationships (Fig. 3B).

A biplot of the first two dimensions (principal components) generated from the genomic relationship matrix revealed that they accounted for 3.5% and 2.9% of the

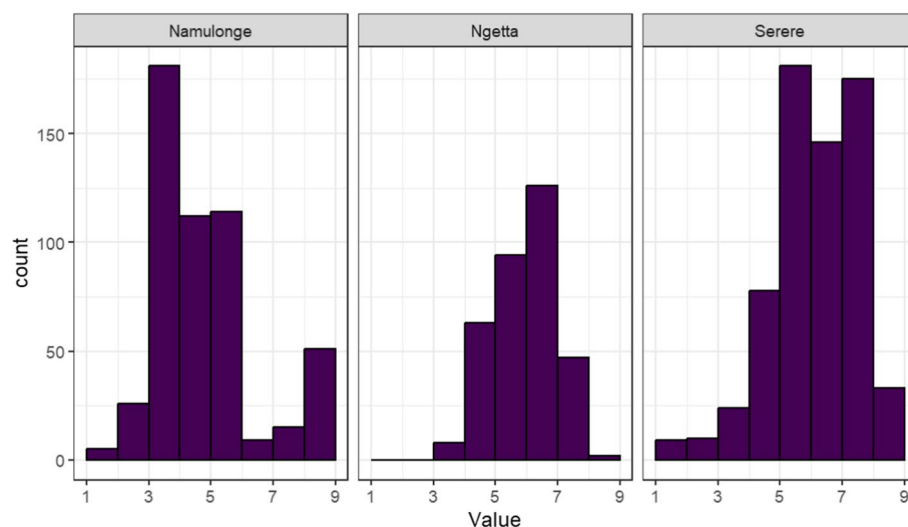


Fig. 2 Distribution of fresh cassava root HCN content across study locations

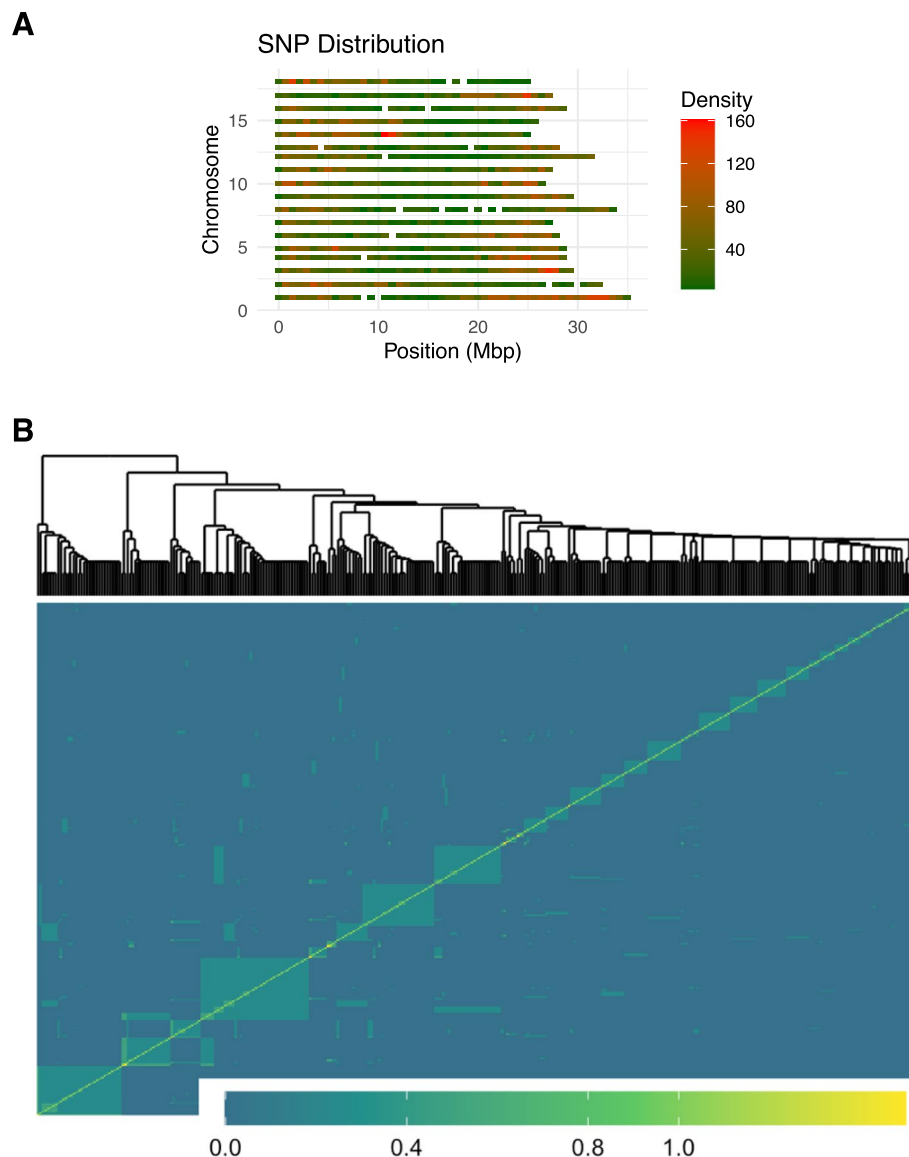


Fig. 3 Distribution of SNPs on chromosomes (A) and heatmap of pairwise genomic relationship matrix based on 19,446 SNPs (B)

variation in SNPs. The first ten principal components accounted for 20.2% of the variation. Clustering is visible on the PCA bi-plot, implying presence of groups of individuals that are closely genetically related within the clusters but genetically distant from others in other clusters (Fig. 4).

Cross validation accuracy and predictive ability for the GBLUP model

The iterations 1, 2 and 4 had positive values for accuracy (0.02–0.58) and predictive ability (0.02–0.57). On the hand, folds 3 and 5 had variable prediction accuracy and prediction ability values, ranging from negative to positive, with the highest variability being recorded in fold 5 where accuracy ranged from –0.41 to 0.68 while predictive ability ranged from –0.4 to 0.68 (Fig. 6). In general, prediction accuracy was low to moderate across folds,

ranging from –0.41 to 0.68 with a mean of 0.22. An identical trend was recorded for the predictive ability across iterations (mean = 0.22) (Fig. 5). The narrow-sense heritability (h^2) was 0.35.

The accuracy of the GBLUP model ranged from –0.41 to 0.48 with an average of 0.22. The predictive ability of the model on the other hand ranged from –0.4 to 0.68, with a mean of 0.2 (Fig. 5). Across iterations, increasing the accuracy increased the predictive ability of the model. The narrow-sense heritability (h^2) was 0.35.

Performance of Bayesian and RKHS models

Narrow-sense heritability (h^2) estimates for the Bayesian models and RKHS model

With the exception of BRR, the narrow-sense heritability estimates were low to moderate for the Bayesian and

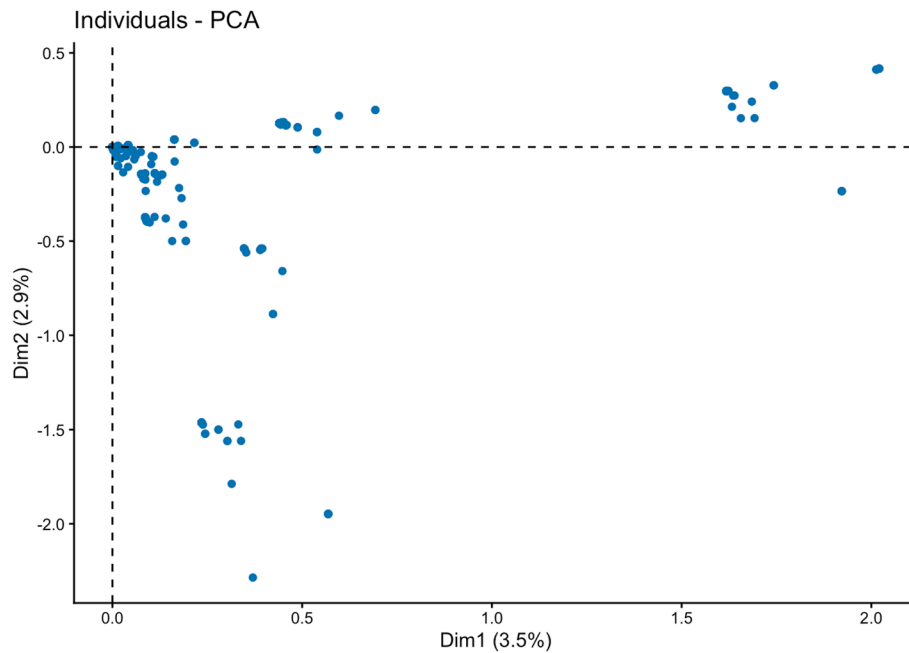


Fig. 4 Principal component analysis biplot derived from the genomic relationship matrix

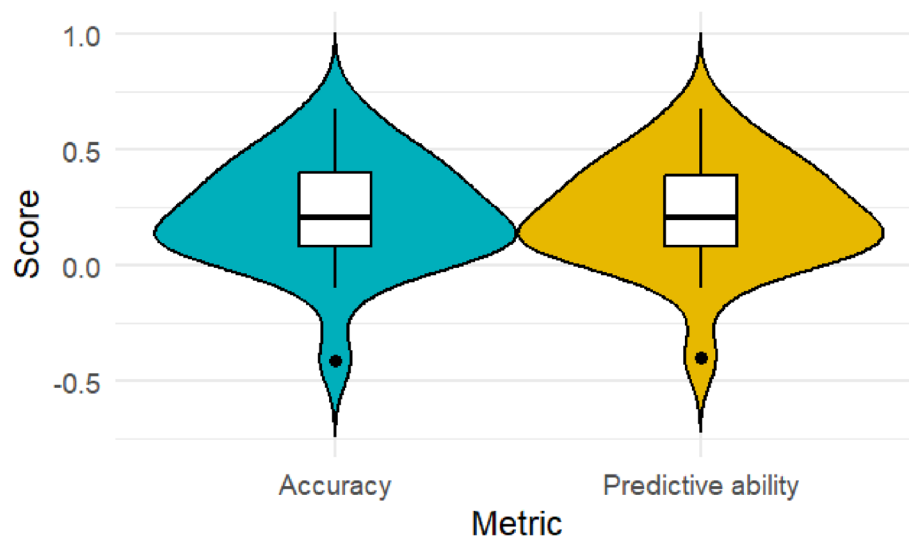


Fig. 5 Variability in accuracy and predictive ability for the GBLUP model

RKHS models across all folds. For Bayes A, h^2 ranged from 0.05 to 0.30, while for Bayes B, it ranged from 0.002 to 0.15. Bayes C showed a range of 0.01 to 0.20, BL ranged from 0.05 to 0.15, and RKHS ranged from 0.13 to 0.18. The highest heritability per fold (0.66 to 0.70) was observed with the BRR model (Fig. 6).

Accuracy of RKHS and Bayesian models

With the exception of BRR, the accuracy of the models was moderate, ranging from 0.48 to 0.52. The Reproducing Kernel Hilbert Spaces and Bayesian Lasso models predicted better ($r=0.52$), followed by the Bayes A model

($r=0.49$) and Bayes B and C models ($r=0.48$). The Bayesian Ridge Regression (BRR) model had a low accuracy ($r=0.18$) (Fig. 7).

Cross validation predictive ability of RKHS and Bayesian models

Generally, the predictive ability of the models across folds was low, ranging from 0.12 to 0.22. Bayes B model had the least PA ($r=0.12$) followed by Bayes C ($r=0.13$), Bayesian Ridge Regression ($r=0.15$), Bayesian Lasso ($r=0.16$) and Bayes A ($r=0.19$). Reproducing Kernel Hilbert Spaces model had higher PA ($r=0.22$) than all the

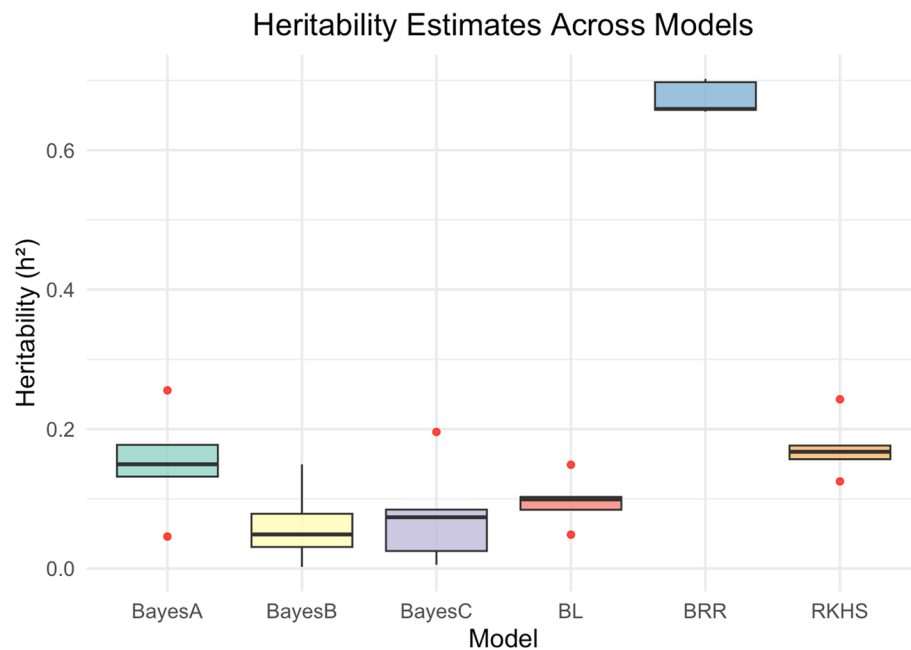


Fig. 6 Box plots showing variability in narrow-sense heritability (h^2) between the Bayesian models and the RKHS model

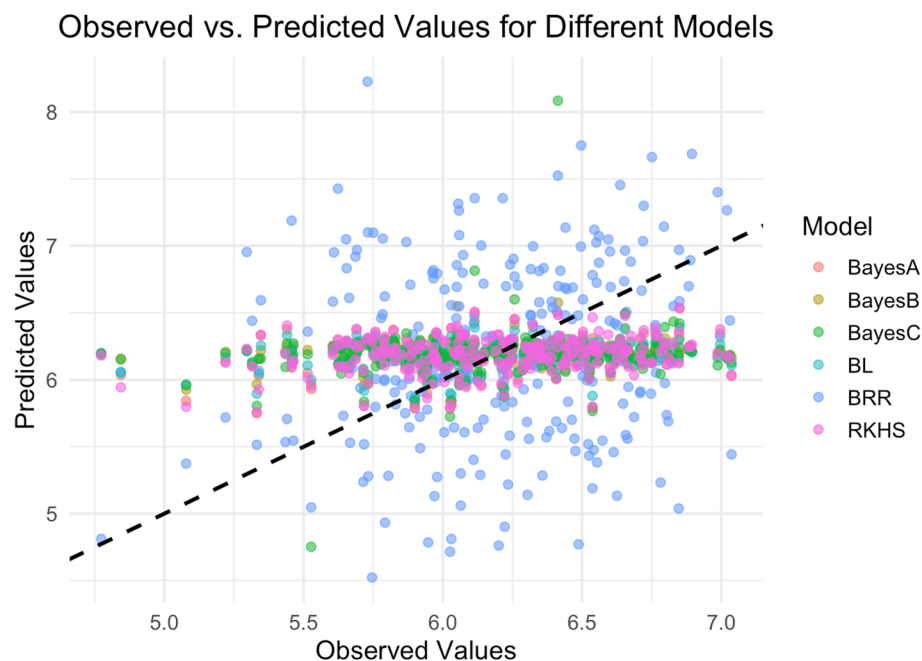


Fig. 7 Scatter plots of observed (actual) HCN phenotype against the predicted HCN phenotype (BLUEs). The black line is the 1:1 line, a reference line where predicted values equal observed values for each model

Bayesian models. All the Bayesian models did not perform differently (at $\alpha = 0.05$) from each other (Fig. 8).

Discussion and conclusions

Cassava is clonally propagated, has a long breeding cycle and complex genetics owing to its out crossing nature. Thus, genetic gains tend to be small compared to other crops [49–51]. Phenotyping for HCN trait is

slow, laborious and may require specialized tools and/or personnel, making it restrictive for most resource constrained breeding programs. Moreover, it can only be done after the crop has attained physiological maturity, twelve months after planting. In contrast, genomic selection relies on genome-wide markers to predict the breeding values of individuals and has demonstrated significant potential, especially in early-stage breeding

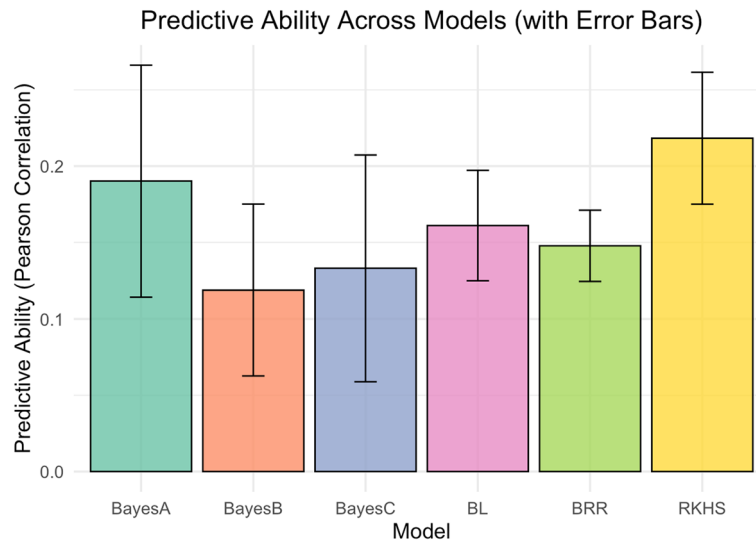


Fig. 8 Variability in predictive ability across Bayesian models and RKHS model

or in scenarios where phenotypic selection is difficult [52], such as with HCN. This study assessed the accuracy and predictive ability of seven genomic selection models—GBLUP, Bayes A, Bayes B, Bayes C, BL, BRR, and RKHS—for predicting the hydrogen cyanide (HCN) content in fresh cassava roots.

Phenotypic variation for HCN ranged from 1 to 9 in Namulonge and Serere and 4 to 9 in Ngetta. This shows that there was high variability for HCN within the study population. This is especially true given the fact that selection for low HCN clones had never been performed in this cyclic population. This variability can partly be explained by two possibilities. First, owing to the heterozygous nature of cassava and the two rounds of recombination to generate cycle 2 populations, it is anticipated that this population could have considerable standing genetic variation. Second, original founders for cycle 2 were sourced from diverse origins (i.e., in Africa and Latin America) and without significant positive selection for HCN. Thus, standing genetic variation was high as reflected in the PCA biplot (Fig. 4). This variability is suggestive of population stratification, which was accounted for using the kinship matrix. Estaghirou et al. [36] demonstrated that genomic prediction accuracy in plant breeding increases with increase in genetic variation, especially when combined with increase in marker density and population size.

Broad-sense heritability and narrow-sense heritability estimates

The broad-sense heritability estimate ($H^2=0.42$) for the study was relatively low but still higher than the narrow-sense heritability (h^2) estimates ($h^2=0.05$ – 0.35). While narrow-sense heritability only considers additive genetic effects, broad-sense heritability captures the total genetic

variation, including additive, non-additive, dominance, and epistatic effects [53]. The h^2 was generally low, suggesting that non-additive effects may play a significant role in the genetic regulation of the trait. However, when using the BRR model, narrow-sense heritability ($h^2=0.66$ – 0.70) surpassed broad-sense heritability, indicating the potential over estimation of additive variance. The BRR model assumes that all markers contribute small, normally distributed effects, thus applies uniform shrinkage [31]. However, since HCN is oligogenic, being controlled by a few large effect loci [20] it is possible that the model could have overestimated the effects of the small-effect loci and is thus over fitted.

Population structure

Population structure, if unaccounted for can artificially inflate heritability estimates and genomic prediction accuracies as the model might capture family specific-signals rather than the true genetic signal [37, 54]. From the heatmap of the genomic relationship matrix (GRM) (Fig. 3B) and the PCA bi-plot (Fig. 4), it is clear that there was structure within the population, occasioned by relatedness among individuals from the same family. This clustering has practical implications for the breeding program. In order for the breeders to optimally utilize the diversity within this population, parents selected for the next cycle of genomic selection should be representative of individuals with superior GEBVs from the different clusters. The mating design selected should also maximize cross combinations between clusters.

To account for the evident population structure, in the GBLUP model, we incorporated the genomic relationship matrix (GRM). The GRM captures genetic similarities and differences between individuals based on SNPs and allows to adjust for any stratification (sub-populations

that might have different allele frequencies) within the population by incorporating them into the genomic prediction model [55]. For the Bayesian models, population structure was accounted for using Markov chain Monte Carlo sampling which integrates population structure by sampling from the posterior distributions [56] while in the RKHS algorithm, population structure was accounted for by fitting the Gaussian kernel which can model complex genetic relationships and interactions within the population [57].

Predictive accuracy and predictive ability of genomic models for HCN

Genomic predictive accuracy and predictive ability are two related but different concepts. Genomic prediction accuracy is the correlation between the genomic estimated breeding values (GEBVs) and the true breeding value divided by the square root of h^2 [58, 59]. Accuracy is therefore a measure of how close predicted values (GEBVs) are to the actual HCN phenotype. Predictive ability (PA) on the other hand reflects how well a model can predict GEBVs for samples not included in the training set, making it a reliable indicator of the model's generalization capability [60].

In this study, PA values were consistently low across all models, ranging from 0.12 to 0.22. Among the models tested, RKHS and GBLUP achieved the highest PA, both reaching 0.22. This is comparable to the PA value (0.26) reported by Torres et al. [25] for HCN in a Nigerian cassava population, but much lower than that reported for Embrapa (0.59) and combined Nigerian and Embrapa populations (0.64) in the same study. For accuracy of genomic predictions, the GBLUP model has previously been demonstrated to be superior or perform as well as other more complex models [19, 28, 61, 62] for other traits while using less computational resources. The GBLUP model takes into consideration genetic relationships between individuals by considering the genomic relation matrix (GRM) constructed from SNP marker information. This contributes to increasing selection gains as covariance information amongst individuals is used to get GEBVs [61]. The prediction accuracy of GBLUP ($r=0.22$) was lower than all other models except BRR ($r=0.18$) and also lower than that reported by Torres et al. [25] for HCN using IITA-Nigerian germplasm ($r=0.51$), Embrapa germplasm ($r=0.67$) and the combined Embrapa and IITA population ($r=0.8-0.85$). However, these predictive accuracies are comparable to accuracies reported for RKHS, Bayes A, Bayes B, Bayes C and BL ($r=0.48-0.52$) by the same authors. The low predictive accuracy reported in this study could be attributed to the low broad-sense heritability of HCN (0.42) in the study population as compared to H^2 of 0.76 and 0.56 reported for Embrapa and Embrapa-IITA populations

[25]. The low H^2 in this study suggests that a large proportion of the observed phenotypic variance in HCN is due non-additive genetic effects. Genomic prediction accuracy has been shown to increase with increase in trait heritability [40, 59].

The RKHS model which was more accurate than GBLUP with equal generalizing ability confers more advantages for selection. Reproducing Kernel Hilbert Spaces Regression was one of the earliest machine learning methods used in plant and animal breeding [63]. Kernel methods have gained wide acceptance due to their ability to effectively capture non-linear patterns in data, which linear statistical methods often fail to address [40]. This has been demonstrated in various studies, including those on cassava [28], chicken [64], and wheat [65, 66] despite the underlying differences in genetic architecture of the traits. Additionally, the RKHS algorithm is capable of accounting for both additive and non-additive effects [40]. However, some studies have found only minor differences between RKHS and linear methods [67, 68], reinforcing the idea that no single genomic prediction model is universally optimal for all datasets, as each model operates under different assumptions which may not hold true for certain traits. The use of multiple genomic prediction models combined into an ensemble has been demonstrated to improve genomic prediction accuracy and reduce prediction errors over individual genomic prediction models in a nested association mapping dataset of teosinte [69]. This as an approach that cassava breeding programs targeting low HCN varieties could also explore.

The superior performance of RKHS and Bayesian models to GBLUP could be attributed to the genetic architecture of HCN and the different assumptions made by the models on the gene effects. The genetic control of HCN is oligogenic, being mainly controlled by two genes on chromosome 14 and chromosome 16 [20]. Bayes A and B assume that marker effects are independent of each other, thus only a few markers are assumed to have large effects on the trait. Bayes C on the other hand assumes that only a small proportion of markers have moderate to large effects while BL assumes that most marker effects are small (zero) but a few may be large, thus unevenly shrinks markers, more aggressively shrinking small effect markers [31, 32]. All these model assumptions are congruent with the genetic architecture of HCN as an oligogenic trait. The superior predictive accuracy of RKHS can be attributed to its ability to model non-additive gene effects (dominance and epistasis) without having to explicitly specifying them in the model given its use of kernels [70]. Dominance and epistatic effects have both been reported for HCN [20]. On the other hand, the GBLUP assumes equal contribution (small effects) for all the markers across the genome [31, 40] thus

fails to capture the enormous contribution of the large effect markers. An alternative to GBLUP would be the weighted GBLUP (WssBLUP) [38] in which markers can be assigned different weights based on their effects. For oligogenic HCN, the GWAS derived significant markers reported by Ogbonna et al. [20] can be incorporated as fixed effects into the GBLUP model such that the model can assign greater weight to these loci while still capturing background polygenic effects.

A major limitation in this study is that ordinal HCN scores were treated like Gaussian distributed quantitative data. For a model like the GBLUP that relies on partitioning of genetic variance following Gaussian assumptions, the use of ordinal data was probably a mismatch between data type and model assumptions and could potentially have contributed to the low prediction accuracy reported with the GBLUP. The use of data generated from the modified picrate paper method which involves using the picrate paper in combination with a spectrophotometer [15] could therefore yield more compelling results.

Practical implications for low cyanide cassava breeding

A recent study [20] reported that the genetic architecture of HCN is highly conserved and identified two major loci encoding for an ATPase and MATE protein, explaining up to 7% and 30% of the HCN phenotypic variation in the cassava root respectively. These loci were also shown to be in strong linkage disequilibrium (LD), which ensures that marker – trait associations are stable. The mate family proteins are involved in detoxification and are capable of transportation of cyanogenic glucosides [71]. This strong LD contributes to the prediction accuracies reported in this study since, proving that genomic prediction models can capture the genetic signal even when under the control of a few markers.

Although this study reports a low to moderate predictive accuracy for HCN, genomic selection could still significantly expedite the recurrent selection cycle by enabling the early selection of parents [22]. Besides recurrent selection, getting estimates of HCN (within limits) on unevaluated plots in multi-locational frameworks/sparse testing networks also enhances selection decisions. This approach could enhance genetic gains per unit of cost and time by eliminating poorly performing individuals, thereby reducing expenses related to field maintenance and phenotyping [72]. Furthermore, genomic selection-driven cyclic population improvement has the potential to swiftly increase the frequency of desirable alleles within a population. For instance, [73] demonstrated that conducting two genomic selection cycles over two years elevated the frequency of African cassava mosaic resistance alleles from 44 to 63%, achieving over twice the gains of traditional recurrent selection, with moderate genomic prediction accuracies of

0.53–0.58. Similarly, [74] observed 4 to 43% higher maize grain yield after two genomic selection cycles compared to pedigree-based selection in bi-parental populations, while [75] reported annual yield improvements of 1.2% to 2.8% in a multi-parental tropical maize population using genomic selection.

Genomic selection improves breeding program efficiency by allowing large populations to be screened early at minimal cost, which not only increases the selection intensity but also reduces the cycle time. For example, under traditional recurrent selection, HCN phenotypic evaluation and subsequent selection is done at crop maturity, which is 12 months after planting. Using genomic selection, genotyping can be done of one month-old seedlings and selection of parents done using GEBVs. This cuts the cycle by over 8 months. This, combined with its ability to increase selection intensity and accuracy presents genomic selection as a suitable vehicle to expeditiously deliver desired genetic gains for low HCN cassava varieties to farmers' fields.

From the results presented, it is evident that choice of model has a big implication on the accuracy of genomic predictions for HCN. It was also apparent that increasing accuracy of the models resulted into improved predictive ability, making the models more reliable. Therefore, it pays for breeders to improve genomic prediction accuracy. Other important considerations for improving HCN prediction accuracy will be improving field experimental designs and optimizing laboratory HCN data collection protocols to minimize the error variance and thus increase trait heritability. Increasing the size of the training population, increasing marker density as well as including HCN trait associated markers are all options that could be explored [28, 38, 76]. The genomic prediction accuracies reported paint a picture of great potential. However, fresh cassava root HCN phenotypic expression is highly dependent on the environment [77]. Indeed, this study has also reported variability for HCN across test environments with Namulonge (mean HCN = 5.3) registering lower average scores compared to Ngetta (mean HCN = 6.4) and Serere (mean HCN = 6.6). In fact, HCN phenotypic expression is so variable that there are significant differences within roots of the same plant, even for different parts of the same root [78]. Thus, breeding programs embarking on HCN phenotyping must adopt standardized protocols for root sampling and processing [42]. Given this environment driven variability in HCN phenotypic expression, at the late stages of breeding when accession numbers are few (advanced yield trials or pre-commercial release stage), confirmatory laboratory based HCN phenotyping is necessary as a quality control step to genomic selection.

Given the phenotypic plasticity of fresh cassava root HCN content, it would be ideal for public sector cassava

breeding programs to aim for low HCN varieties that are adapted to specific environments. However, given the chronic resource constraints, most programs opt to breed for wide adaptability by testing in representative target environments. To optimize resources, while deploying genomic selection, it would be ideal to use data of one environment to predict the performance of clones in another environment. However, when the genotype by environment signal is strong as is the case for HCN [77], the genetic signal captured in one environment may not be transferrable to another [79, 80]. Thus, the use of multi-environment data and genomic selection models that account for the genotype by environment interaction helps to improve the utility of the predictions. Ultimately, breeding for low HCN cassava varieties not only improves food safety but is also a gender responsive intervention that promises to relieve women of the burden of manual processing to detoxify cassava which disproportionately falls on them [81].

Although the cost of genotyping and thus conducting genomic selection has drastically reduced over the years [82], an estimated \$21 per sample for high density genotyping with DArTseq [83] is still very high for publicly funded breeding programs especially given that they typically evaluate hundreds to thousands of accessions annually. Thus, there is need for cheaper and equally accurate or even more accurate methods for predicting performance and facilitating selection of desired individuals in large breeding programs [84] proposed the use of near infrared spectroscopy (NIRS) as a high throughput, low cost and non-destructive tool for indirectly capturing endophenotype variation and computing relationship matrices for predicting complex traits. This was termed as phenomic prediction (PP).

While GS relies on genome-wide markers to predict the phenotype of individuals, PP replaces molecular markers with NIRS spectra to deduce relationships among individuals and do selection, attaining results comparable to GS [85]. The SNP marker derived genomic relationship matrix is substituted with a NIRS spectra derived relationship matrix [85]. Since NIRS spectra reveal the biochemical composition of tissue, the variation observed in these spectra is linked to the chemical bonds within the sample tissue (endophenotypes). The principle behind NIRS aligns with the Beer-Lambert law [86], which states that absorbance or reflectance at a specific wavelength is directly proportional to the analyte's concentration. Reflectance at each wavelength represents an integration of various endophenotypic variations, meaning that a relationship matrix built from these spectra could potentially capture a genetic signal [84, 87]. Moreover, spectra could be used simultaneously with molecular markers to increase genomic prediction accuracies [88, 89].

In conclusion, our study has provided a critical evaluation of genomic prediction models for HCN in fresh cassava roots and demonstrated that genomic prediction can be effectively used to breed for low HCN cassava varieties. Whereas the commonly used GBLUP model showed low accuracy, the RKHS and the Bayesian Lasso model out performed all other models, achieving moderate accuracy. This highlights the importance of selecting models that capture both the additive and non-additive genetic effects when making genomic predictions for fresh cassava root HCN content. This work lays important groundwork for cassava breeding programs to integrate genomic selection for low HCN in their variety development pipeline, thus accelerating the rate of delivering desired genetic gains to farmers' fields.

Acknowledgements

The authors are grateful to staff of the Nutrition and Bioanalytical Laboratory at NaCRRI for their dedicated work during HCN phenotyping and to the field staff and management of NaCRRI, the National Semi-arid Resources Research Institute and Ngetta Zonal Agricultural Research Institute for hosting the field trials.

Authors' contributions

MK-Conceptualization, data collection, data analysis, writing original draft. MSB-Conceptualization, supervision, reviewing. EN—Conceptualization, supervision, reviewing and editing. ISK – Conceptualization, data curation, data analysis, reviewing and editing. HW- data collection, reviewing and editing. PI- trial management, reviewing and editing. JKB – data collection, reviewing and editing. NM – Trial management, data collection, review and editing. ARN – data collection, review and editing. EW – Trial management, data collection, review and editing. WE – reviewing and editing. RK- Conceptualization, funding acquisition, supervision, project management, data collection, review and editing. All authors read and approved the final manuscript.

Funding

This work was supported by the Nextgen Cassava Breeding Project through a grant to Cornell University by Bill & Melinda Gates Foundation (BMGF) (Grant INV-007637) and the UK's Foreign Common Wealth and Development Office (FCDO). Additional funding was provided by the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) through grant RU-NARO/2020/post-Doc/01 to NaCRRI.

Data availability

Data sets generated and/or analysed are available from Cassavabase, an open access cassava breeding database via the link the [<https://cassavabase.org/breeders/trial/6707?format=>].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 8 November 2025 / Accepted: 25 March 2026

Published online: 31 March 2026

References

- Okogbenin E, Setter TL, Ferguson M, Mutegi R, Ceballos H, Olanmi B, et al. Phenotypic approaches to drought in cassava: Review. *Front Physiol*. 2013;4(MAY(May)):1–15. <https://doi.org/10.3389/fphys.2013.00093>.
- Burns A, Gleadow R, Cliff J, Zacarias A, Cavagnaro T. Cassava: the drought, war and famine crop in a changing world. *Sustainability*. 2010;2(11). <https://doi.org/10.3390/su2113572>.
- Zagrobeleny M, Bak S, Rasmussen AV, Jørgensen B, Naumann CM, Møller BL. Cyanogenic glucosides and plant-insect interactions. *Phytochemistry*. 2004;65(3):293–306. <https://doi.org/10.1016/j.phytochem.2003.10.016>. PubMed PMID: 14751300.
- Jørgensen K, Bak S, Busk PK, Sørensen C, Olsen CE, Puonti-Kaerlas J, et al. Cassava plants with a depleted cyanogenic glucoside content in leaves and tubers. Distribution of cyanogenic glucosides, their site of synthesis and transport, and blockage of the biosynthesis by RNA interference technology. *Plant Physiol*. 2005. pp. 363–74. <https://doi.org/10.1104/pp.105.065904>. PubMed PMID: 16126856.
- Chiwona-Karlun L, Brimer L, Kalenga Saka JD, Mhone AR, Mkumbira J, Johansson L, et al. Bitter taste in cassava roots correlates with cyanogenic glucoside levels.pdf. *J Sci Food Agric*. 2004;54:581–90. <https://doi.org/10.1002/jsfa.1699>.
- Imakumbili MLE, Semu E, Semoka JMR, Abass A, Mkamilo G. Soil nutrient adequacy for optimal cassava growth, implications on cyanogenic glucoside production: a case of konzo-affected Mtwara region, Tanzania. *PLoS One*. 2019;14(5):1–17. <https://doi.org/10.1371/journal.pone.0216708>. PubMed PMID: 31083702.
- Kashala-Abotnes E, Okitundu D, Mumba D, Boivin MJ, Tylleskär T, Tshala-Katumbay D. Konzo: a distinct neurological disease associated with food (cassava) cyanogenic poisoning. *Brain Res Bull*. 2019;145:87–91. <https://doi.org/10.1016/j.brainresbull.2018.07.001>. PubMed PMID: 29981837.
- Cassava Cyanide Diseases & Neurolathyrism Network (CCDN). Working together to eliminate cyanide poisoning, konzo, tropic ataxic neuropathy (TAN) and neurolathyrism. ISSN 1838-. 2011:1–4. Available from: <https://biblio.ugent.be/publication/2002992/file/2003018.pdf>. (ISSN 1838–8817 (Print): ISSN 1838–8825 (Online)).
- Kinshasa I, Congo DR, Congo DR, Lambein F. Cassava Cyanide Diseases & Neurolathyrism. Network. 2014;8817(24):1–12.
- Alitubeera PH, Eyu PK, Benon A, Alex R, Zhu B ping. Outbreak of cyanide poisoning caused by consumption of cassava flour. Vol. 68. 2019. Report. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6611475/>.
- Atim BP, Kabami Z, Gidudu S. Cyanide Poisoning Investigation, Terego District, Uganda, February 2023. Uganda Natl Inst Public Health Q Epidemiol Bull. 2023;8(2):e5030–41. Available from: <https://uniph.go.ug/cyanide-poisoning-in-vestigation-terego-district-uganda-february-2023/>. Cited 2024 Nov 9.
- Akintonwa A, Tunwashe OL. Fatal cyanide poisoning from cassava-based meal. *Hum Exp Toxicol*. 1992;11:47–9.
- Egan SV, Yeoh HH, Bradbury JH. Simple picrate paper kit for determination of the cyanogenic potential of cassava flour. *J Sci Food Agric*. 1998;76(1):39–48. [https://doi.org/10.1002/\(sici\)1097-0010\(199801\)76:1%3c39::aid-jsfa947%3e3.3.co;2-d](https://doi.org/10.1002/(sici)1097-0010(199801)76:1%3c39::aid-jsfa947%3e3.3.co;2-d).
- Haque MR, Bradbury JH. Total cyanide determination of plants and foods using the picrate and acid hydrolysis methods. *Food Chem*. 2002;77:107–14. [https://doi.org/10.1016/S0308-8146\(01\)00313-2](https://doi.org/10.1016/S0308-8146(01)00313-2).
- Bradbury MG, Egan SV, Bradbury JH. Picrate paper kits for determination of total cyanogens in cassava roots and all forms of cyanogens in cassava products. *J Sci Food Agric*. 1999;79(4):593–601. [https://doi.org/10.1002/\(SICI\)1097-0010\(19990315\)79:4<593::AID-JSFA222/3E3.0.CO;2-2](https://doi.org/10.1002/(SICI)1097-0010(19990315)79:4<593::AID-JSFA222/3E3.0.CO;2-2).
- Bradbury JH, Egan SV, Lynch MJ. Analysis of cyanide in cassava using acid hydrolysis of cyanogenic glucosides. *J Sci Food Agric*. 1991;55(2):277–90. <https://doi.org/10.1002/jsfa.2740550213>.
- Ceballos H, Rojanaridpiched C, Phumichai C. Excellence in cassava breeding: perspectives for the future. *Crop Breed Genet Genom*. 2020;2(2):1–31. <https://doi.org/10.20900/cbpg20200008>.
- Esuma W, Ozimati A, Kulakow P, Gore MA, Nuwamanya E, Egesi C, et al. Effectiveness of genomic selection for improving provitamin A carotenoid content and associated traits in cassava. *G3 Genes|Genomes|Genetics*. 2021;11(May). <https://doi.org/10.1093/g3journal/jkab160>.
- Ozimati AA, Esuma W, Manze F, Iragaba P, Kanaabi M, Ano CU, et al. Utility of Ugandan genomic selection cassava breeding populations for prediction of cassava viral disease resistance and yield in West African clones. *Front Plant Sci*. 2022;13(November). <https://doi.org/10.3389/fpls.2022.1018156>.
- Ogbonna AC, de Andrade BLR, Rabbi IY, Mueller LA, de Oliveira JE, Bauchet GJ. Large-scale genome-wide association study, using historical data, identifies conserved genetic architecture of cyanogenic glucoside content in cassava (*Manihot esculenta* Crantz) root. *Plant J*. 2021;105(3):754–70. <https://doi.org/10.1111/tpj.15071>. PubMed PMID: 33164279.
- Arruda MP, Lipka AE, Brown PJ, Krill AM, Thurber C, Brown-Guedira G, et al. Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). *Mol Breeding*. 2016;36(7):1–11. <https://doi.org/10.1007/S11032-016-0508-5/METRICS>.
- Heffner EL, Jannink JL, Iwata H, Souza E, Sorrells ME. Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci*. 2011;51(6):2597–606. <https://doi.org/10.2135/cropsci2011.05.0253>.
- Kizito Balyejusa Elizabeth, Egwang T, Gullberg U, Fregene M, Westerbergh A. Quantitative trait loci controlling cyanogenic glucoside and dry matter content in cassava (*Manihot esculenta* Crantz) roots. 2007;136:129–36. <https://doi.org/10.1111/j.2007.0018-0661.01975.x>.
- Whankaew S, Poopear S, Kanjanawattana S, Tangphatsornruang S, Boonseng O, Lightfoot DA, et al. A genome scan for quantitative trait loci affecting cyanogenic potential of cassava root in an outbred population. *BMC Genomics*. 2011;12:266. <https://doi.org/10.1186/1471-2164-12-266>.
- Torres LG, Oliveira EJ De, Ogbonna AC, Fonseca F, Simiqueli GF, Kantar MB. Can cross-country genomic predictions be a reasonable strategy to support germplasm exchange ? – A Case Study With Hydrogen Cyanide in Cassava. *Front Plant Sci*. 2021;12(December). <https://doi.org/10.3389/fpls.2021.742638>.
- Kanaabi M, Mukasa SB, Nuwamanya E, Iragaba P, Baguma JK, Nanyonjo AR, et al. Validation of KASP markers associated with hydrogen cyanide in fresh cassava roots in Uganda Cassava Germplasm. *Agronomy*. 2024;14(12):2765. <https://doi.org/10.3390/agronomy14122765>.
- Ozimati A, Kawuki R, Esuma W, Kayondo IS, Wolfe M. Training population optimization for prediction of cassava brown streak disease resistance in West African Clones. *G3:Genes|Genomes|Genetics*. 2018;12(8):3903–13. <https://doi.org/10.1534/g3.118.200710>.
- Kayondo SI, Pino D, Carpio D, Lozano R, Ozimati A, Yona B, et al. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. *Sci Rep*. 2018;8:1549. <https://doi.org/10.1038/s41598-018-19696-1>.
- VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci*. 2008;91(11):4414–23. <https://doi.org/10.3168/JDS.2007-0980>. PubMed PMID: 18946147.
- Ilka JJ, Meuwissen THE, Kranis A, Woolliams JA. Use and optimization of different sources of information for genomic prediction. *Genet Sel Evol*. 2017;49(1):1–11. <https://doi.org/10.1186/S12711-017-0365-7/FIGURES/6>. PubMed PMID: 29228899.
- Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157(4):1819–29. <https://doi.org/10.1093/genetics/157.4.1819>. PubMed PMID: 11290733.
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. 2011;12(1):1–12. <https://doi.org/10.1186/1471-2105-12-186/FIGURES/2>. PubMed PMID: 21605355.
- Park T, Casella G. The Bayesian lasso. *J Am Stat Assoc*. 2008;103(482):681–6. <https://doi.org/10.1198/016214508000000337>.
- Barili LD, Vale NM do, Silva FF e, Carneiro JE de S, Oliveira HR de, Vianello RP, et al. Genome prediction accuracy of common bean via Bayesian models. *Ciência Rural*. 2018;48(8). <https://doi.org/10.1590/0103-8478cr20170497>.
- Wolpert DH. The lack of a priori distinctions between learning algorithms. *Neural Comput*. 1996;8:1341–90. <https://doi.org/10.1162/neco.1996.8.7.1341>.
- Estaghirou SBO, Ogutu JO, Piepho HP. How genetic variance and number of genotypes and markers influence estimates of genomic prediction accuracy in plant breeding. *Crop Sci*. 2015;55(5):1911–24. <https://doi.org/10.2135/CROPSCI2014.09.0620>.
- Werner CR, Gaynor RC, Gorjanc G, Hickey JM, Kox T, Abbadi A, et al. How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. *Front Plant Sci*. 2020;11. <https://doi.org/10.3389/fpls.2020.592977>.
- Pang Z, Wang W, Huang P, Zhang H, Zhang S, Yang P, et al. Enhancing genomic prediction accuracy with a single-step genomic best linear unbiased prediction model integrating genome-wide association study results. *Animals*. 2025;15(9):1268. <https://doi.org/10.3390/ANI15091268>.
- Fukuda WMG, Guevara CL, Kawuki R, Ferguson ME. Selected morphological and agronomic descriptors for the characterization of cassava. IITA, Ibadan, Nigeria. 2010. Report. Available from: <https://books.google.co.ug/books?hl=en&lr=&id=-SnckHhBIEYC&oi=fnd&pg=PA1&dq=FUKUDA+2010+hydrogen+c>

- yanide+cassava+scale&ots=_tpYNzi4sg&sig=QbPKY4TOMSnmZK6biD6Z7XztjUQ&redir_esc=y#v=onepage&q&f=false. Cited 2022 Mar 29.
40. Montesinos López OA, Montesinos López A, Crossa J. Multivariate statistical machine learning methods for genomic prediction. *Multivariate statistical machine learning methods for genomic prediction*. 2022. <https://doi.org/10.1007/978-3-030-89010-0>.
 41. Lourenço VM, Ogutu JO, Rodrigues RAP, Posekany A, Piepho HP. Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. *BMC Genomics*. 2024;25(1):1–20. <https://doi.org/10.1186/S12864-023-09933-X>. PubMed PMID: 38326768.
 42. Michael, Kanaabi, Ephraim Nuwamanya KM and TT. NIRS Acquisition on Fresh Cassava Roots Using FOSS DS2500 and ASD Quality Spec and Relating the NIRS Spectra to Fresh HCN Content by Picrate Method. CIRAD data verse. 2022. p. 11.
 43. Gezan SA, de Oliveira Giovanni Galli AA, Murray D, Oliveira de, Gezan salvadoregezan SA, Amanda de Oliveira vsnicouk A, et al. ASRgenomics: An R package with Complementary Genomic Functions. Version 1.1.0. VSN International. 2022. Report. Available from: <https://vsn.co.uk/free-software/asrgenomics>.
 44. Spilke J, Piepho HP, Hu X. A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *J Agric Biol Environ Stat*. 2005;10(3):374–89. <https://doi.org/10.1198/108571105X58199/METRICS>.
 45. Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R. ASReml-R Reference Manual Version 4.2 ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML). 2023. Report. Available from: <https://asreml.kb.vsnr.co.uk/article-categories/asreml-r-resources/>.
 46. Pérez P, De Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 2014;198(2):483–95. <https://doi.org/10.1534/GENETICS.114.164442/-DC1>. PubMed PMID: 25009151.
 47. Sallam AH, Endelman JB, Jannink JL, Smith KP. Assessing Genomic Selection Prediction Accuracy in a Dynamic Barley Breeding Population. *Plant Genome*. 2015;8(1). <https://doi.org/10.3835/plantgenome2014.05.0020>.
 48. Chen CY, Misztal I, Aguilar I, Tsurtur S, Meuwissen THE, Aggrey SE, et al. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. *J Anim Sci*. 2011;89(1):23–8. <https://doi.org/10.2527/JAS.2010-3071>. PubMed PMID: 20889689.
 49. Ceballos H, Kulakow P, Hershey C. Cassava breeding: current status, bottlenecks and the potential of biotechnology tools. *Trop Plant Biol*. 2012;5(1):73–87. <https://doi.org/10.1007/s12042-012-9094-9>.
 50. Ceballos H, Iglesias CA, Pérez JC, Dixon AGO. Cassava breeding: opportunities and challenges. *Plant Mol Biol*. 2004;56(4):503–16. <https://doi.org/10.1007/S11033-004-5010-5/METRICS>. PubMed PMID: 15630615.
 51. Manze F, Rubaihayo P, Ozimati A, Gibson P, Esuma W, Bua A, et al. Genetic gains for yield and virus disease resistance of cassava varieties developed over the last eight decades in Uganda. *Front Plant Sci*. 2021;12:1–11. <https://doi.org/10.3389/fpls.2021.651992>.
 52. Crossa J, Pérez P, Hickey J, Burguño J, Ornella L, Cerón-Rojas J, et al. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*. 2013;112(1):48–60. <https://doi.org/10.1038/hdy.2013.16>. PubMed PMID: 23572121.
 53. Schmidt P, Hartung J, Bennewitz J, Hans-Peter P. Heritability in plant breeding on a genotype-difference basis. *Genetics*. 2019;212(4):991–1008. <https://doi.org/10.1534/genetics.119.302134>. PubMed PMID: 31248886.
 54. Hayatgheibi H, Hallingbäck HR, Lundqvist SO, Grahn T, Scheepers G, Nordström P, et al. Implications of accounting for marker-based population structure in the quantitative genetic evaluation of genetic parameters related to growth and wood properties in Norway spruce. *BMC Genom Data*. 2024;25(1):1–14. <https://doi.org/10.1186/S12863-024-01241-X/TABLES/6>. PubMed PMID: 38877416.
 55. Raymond B, Wientjes YCJ, Bouwman AC, Schrooten C, Veerkamp RF. A deterministic equation to predict the accuracy of multi-population genomic prediction with multiple genomic relationship matrices. *Genet Sel Evol*. 2020;52(1):1–22. <https://doi.org/10.1186/S12711-020-00540-Y>. PubMed PMID: 32345213.
 56. Wolc A, Dekkers JCM. Application of Bayesian genomic prediction methods to genome-wide association analyses. *Genet Sel Evol*. 2022;54(1):1–12. <https://doi.org/10.1186/S12711-022-00724-8>. PubMed PMID: 35562659.
 57. De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res*. 2010;92(4):295–308. <https://doi.org/10.1017/S0016672310000285>. PubMed PMID: 20943010.
 58. Goddard ME, Hayes BJ. Genomic selection. *J Anim Breed Genet*. 2007;124(6):323–30. <https://doi.org/10.1111/J.1439-0388.2007.00702.X>. PubMed PMID: 18076469.
 59. Shengqiang Z, Dekkers JCM, Fernando RL, Jannink JL. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics*. 2009;182(1):355–64. <https://doi.org/10.1534/GENETICS.108.098277>. PubMed PMID: 19299342.
 60. Faville MJ, Ganesh S, Cao M, Jahufer MZZ, Bilton TP, Easton HS, et al. Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor Appl Genet*. 2018;131(3):703–20. <https://doi.org/10.1007/S00122-017-3030-1/FIGURES/5>. PubMed PMID: 29264625.
 61. de Andrade LRB, Bandeira e Sousa M, Oliveira EJ, de Resende MDV, Azevedo CF. Cassava yield traits predicted by genomic selection methods. *PLoS One*. 2019;14(11):e0224920. <https://doi.org/10.1371/JOURNAL.PONE.0224920>.
 62. Wolfe MD, Del Carpio DP, Alabi O, Ezenwaka LC, Ikeogu UN, Kayondo IS, et al. Prospects for Genomic Selection in Cassava Breeding. *Plant Genome*. 2017;10(3). <https://doi.org/10.3835/plantgenome2017.03.0015>. PubMed PMID: 29293806.
 63. Gianola D, Van Kaam JBCHM. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*. 2008;178(4):2289–303. <https://doi.org/10.1534/genetics.107.084285>. PubMed PMID: 18430950.
 64. Long N, Gianola D, Rosa GJM, Weigel KA, Kranis A, Gonzalez-Recio O. Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genet Res Camb*. 2010;92(3):209–25. <https://doi.org/10.1017/S016672310000157>. PubMed PMID: 20667165.
 65. Crossa J, Perez P, De Los Campos G, Mahuku G, Dreisigacker S, Magorokosho C. Genomic prediction of quantitative traits in plant breeding using molecular markers and pedigree. 2010. Report. Available from: https://www.academi.edu/download/105971958/Chapter_DRAFT.pdf. Cited 2025 Jan 20.
 66. Cuevas J, Montesinos-López O, Juliana P, Guzmán C, Pérez-Rodríguez P, González-Bucio J, et al. Deep Kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 Genes[Genomes]Genetics*. 2019;9(9):2913–24. <https://doi.org/10.1534/g3.119.400493>.
 67. Morota G, Bodhireddy P, Vukasinovic N, Gianola D, DeNise S. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. *Front Genet*. 2014;5(MAR):75304. <https://doi.org/10.3389/FGENE.2014.00056/BIBTEX>.
 68. Tusell L, Pérez-Rodríguez P, Forni S, Wu XL, Gianola D. Genome-enabled methods for predicting litter size in pigs: a comparison. *Animal*. 2013;7(11):1739–49. <https://doi.org/10.1017/S1751731113001389>.
 69. Tomura S, Wilkinson MJ, Cooper M, Powell O. Improved genomic prediction performance with ensembles of diverse models. *G3: Genes, Genomes, Genetics*. 2025;15(5):48. <https://doi.org/10.1093/g3journal/jka048>.
 70. Montesinos-López OA, Carter AH, Bernal-Sandoval DA, Cano-Paez B, Montesinos-López A, Crossa J. A comparison between three tuning strategies for Gaussian kernels in the context of univariate genomic prediction. *Genes (Basel)*. 2022;13(12):2282. <https://doi.org/10.3390/GENES13122282/S1>. PubMed PMID: 36553547.
 71. Zou Z, Yang J. Genome-wide comparison reveals divergence of cassava and rubber aquaporin family genes after the recent whole-genome duplication. *BMC Genomics*. 2019;20(1). <https://doi.org/10.1186/s12864-019-5780-4>. PubMed PMID: 31092186.
 72. Edriss V, Gao Y, Zhang X, Jumbo MB, Makumbi D, Olsen MS, et al. Genomic prediction in a large African maize population. *Crop Sci*. 2017;57(5):2361–71. <https://doi.org/10.2135/CROPSCI2016.08.0715>.
 73. Wolfe MD, Rabbi IY, Egesi C, Hamblin M, Kawuki R, Kulakow P, et al. Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *Plant Genome*. 2016;9(2):1–13. <https://doi.org/10.3835/plantgenome2015.11.0118>.
 74. Vivek BS, Krishna GK, Vengadesan V, Babu R, Zaidi PH, Kha LQ, et al. Use of Genomic Estimated Breeding Values Results in Rapid Genetic Gains for Drought Tolerance in Maize. *Plant Genome*. 2017;10(1). <https://doi.org/10.3835/PLANTGENOME2016.07.0070>. PubMed PMID: 28464061.

75. Zhang A, Wang H, Beyene Y, Semagn K, Liu Y. Effect of Trait Heritability , Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations. *Front Plant Sci*. 2017;8(November):1–12. <https://doi.org/10.3389/fpls.2017.01916>.
76. Somo M, Kulembeka H, Mtunda K, Mrema E, Salum K, Wolfe MD, et al. Genomic prediction and quantitative trait locus discovery in a cassava training population constructed from multiple breeding stages. *Crop Sci*. 2020;60(2):896–913. <https://doi.org/10.1002/CSC2.20003>.
77. Kanaabi M, Settumba MB, Nuwamanya E, Muhumuza N, Iragaba P, Ozimati A, et al. Genetic Variation and Heritability for Hydrogen Cyanide in Fresh Cassava Roots: Implications for Low-Cyanide Cassava Breeding. *Plants*. 2024;13(9). <https://doi.org/10.3390/plants13091186>.
78. Nuwamanya E, Turyasingura C, Magumba I, Katungisa A, Alicai T. Cyano-genic potential variations within plot, plant and roots of cassava varieties grown in the same environment. *Proc Natl Acad Sci India Section B Biol Sci*. 2023;93(2):365–72. <https://doi.org/10.1007/S40011-022-01418-3/METRICS>.
79. Hassen MB, Bartholomé J, Valé G, Cao TV, Ahmadi N. Genomic Prediction Accounting for Genotype by Environment Interaction Offers an Effective Framework for Breeding Simultaneously for Adaptation to an Abiotic Stress and Performance Under Normal Cropping Conditions in Rice. *G3 Genes[Genomes]Genetics*. 2018;8(7):2319–32. <https://doi.org/10.1534/g3.118.200098>. PubMed PMID: 29743189.
80. de Oliveira AA, Resende MFR, Ferrão LFV, Amadeu RR, Guimarães LJM, Guimarães CT, et al. Genomic prediction applied to multiple traits and environments in second season maize hybrids. *Heredity*. 2020;125(1):60–72. <https://doi.org/10.1038/s41437-020-0321-0>. PubMed PMID: 32472060.
81. Chiwona-Karlton L, Katundu C, Ngoma C, Chipungu F, Mkumbira Jonathan, Simukoko S, et al. Bitter cassava and women: an intriguing response to food security. *Leisa Magazine*. 2002;18(4).
82. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, et al. Adaptation genomics : the next generation. *Trends Ecol Evol*. 2010;25(12):705–12. <https://doi.org/10.1016/j.tree.2010.09.002>.
83. Mbanjo EGN, Ogungbesan A, Agbona A, Akpotuzor P, Toyinbo S, Iluebbey P, et al. Validation of SNP markers for diversity analysis, quality control, and trait selection in a biofortified cassava population. *Plants*. 2024;13(16):2328. <https://doi.org/10.3390/plants13162328>.
84. Rincint R, Charpentier JP, Faivre-Rampant P, Paux E, Le Gouis J, Bastien C, et al. Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. *G3 Genes[Genomes]Genetics*. 2018;8(12):3961–72. <https://doi.org/10.1534/G3.118.200760>. PubMed PMID: 30373914.
85. Robert P, Brault C, Rincint R, Segura V. Phenomic selection: a new and efficient alternative to genomic selection. *Methods Mol Biol*. 2022;2467:397–420. https://doi.org/10.1007/978-1-0716-2205-6_14. PubMed PMID: 35451784.
86. Swinehart DF. The Beer-Lambert Law. *J Chem Educ*. 1962;39(7):333–5. <https://doi.org/10.1021/ED039P333>.
87. Brault C, Lazerges J, Doligez A, Thomas M, Ecarnot M, Roumet P, et al. Interest of phenomic prediction as an alternative to genomic prediction in grapevine. *Plant Methods*. 2022;1–14. <https://doi.org/10.1186/s13007-022-00940-9>.
88. Hayes BJ, Panozzo J, Walker CK, Choy AL, Kant S, Wong D, et al. Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theor Appl Genet*. 2017;130(12):2505–19. <https://doi.org/10.1007/S00122-017-2972-7>.
89. Sun J, Rutkoski JE, Poland JA, Crossa J, Jannink JL, Sorrells ME. Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *Plant Genome*. 2017;10(2):plantgenome2016.11.0111. <https://doi.org/10.3835/PLANTGENOME2016.11.0111>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.