Ensemble-based genomic prediction for maize flowering time reveals novel insights into trait genetic architecture and improves prediction for breeding applications

Shunichiro Tomura[1,2,*], Owen Powell[1,2,3], Melanie J. Wilkinson[1,2,3], Mark Cooper[1,2,3]

[1]Queensland Alliance for Agriculture and Food Innovation (QAAFI), Centre for Crop Science, the University of Queensland, St Lucia, QLD 4072, Australia
[2]ARC Centre of Excellence for Plant Success in Nature and Agriculture, the University of Queensland, St Lucia, QLD 4072, Australia
[3]ARC Training Centre in Predictive Breeding, the University of Queensland, St Lucia, QLD 4072, Australia

*Corresponding author: s.tomura@uq.edu.au

ORCID ID
Shunichiro Tomura: 0009-0005-9050-9963
Owen Powell: 0000-0001-5131-4286
Melanie J. Wilkinson: 0000-0003-1400-3176
Mark Cooper: 0000-0002-9418-3359

Abstract

While many genomic prediction models have been evaluated for their potential to accelerate genetic gain for multiple traits, no individual genomic prediction model has outperformed others across all applications. This problem aligns with the implications of the No Free Lunch Theorem, stating that the average performance of individual prediction models becomes equivalent across the state space of diverse prediction scenarios. Ensembles of multiple individual genomic prediction models can be a potential alternative approach. The framework of the Diversity Prediction Theorem suggests the potential for a reduction in prediction error with the inclusion of diverse prediction models in the ensemble. We investigated the performance of an ensemble approach that combines multiple genomic prediction models. We demonstrate the results using flowering time traits measured in two maize Nested Association Mapping datasets. For both datasets, the ensemble-based prediction approach achieved the highest prediction accuracy and lowest prediction error across traits. Multiple genomic regions containing key flowering time-related genes were captured by the different genomic prediction models with diverse weights, demonstrating different views of the trait genetic architecture. The combination of such diverse views contributed to the improvement of prediction performance by the ensemble-based approach over the individual prediction models. Exploiting the expectations of the Diversity

Prediction Theorem, the ensemble can overcome some limitations proposed by the No Free Lunch Theorem when applying individual genomic prediction models.

Key words

Genomic prediction, ensemble, genetic architecture, crop breeding, maize, flowering time

Key message

Applying the Diversity Prediction Theorem, an ensemble-based prediction leveraging multiple individual genomic prediction models improved the prediction performance over the individual models by combining multiple views of trait genetic architecture

Introduction

Crop breeding programs have targeted the development of more resilient genotypes for harsh biotic and abiotic stress conditions (Kholová et al. 2021; Cooper and Messina 2023; Werner et al. 2025). The pressures of climate change have amplified such stresses in agricultural environments, resulting in increasing demand for better-adapted crop genotypes (Ceccarelli et al. 2010; Langridge et al. 2021). Consequently, breeders have endeavoured to continuously select genotypes with improved combinations of the target traits to accelerate genetic gain (Messina et al. 2023; Pixley et al. 2023).

Genomic selection (Meuwissen et al. 2001; Bernardo and Yu 2007; Voss-Fels et al. 2019; Crossa et al. 2025; Escamilla et al. 2025) is a key approach to accelerate genetic gain. The prediction of trait phenotypes using genomic markers can be leveraged to reduce the length of each breeding cycle and therefore the cost of breeding programs (Heffner et al. 2010). Genomic prediction models are trained to recognise predictive patterns in genomic marker datasets to represent the genetic architecture of traits (Alemu et al. 2024). Genomic prediction models that can identify a more complete view of trait genetic architecture are sought to improve the prediction performance (Hammer et al. 2006).

Numerous genomic prediction models have been investigated across crop breeding programs (Heslot et al. 2012; John et al. 2022; Crossa et al. 2025; Escamilla et al. 2025). However, none of the proposed approaches have shown superior performance for all traits and datasets, interpreted to be a consequence of the No Free Lunch Theorem (Wolpert and Macready 1997), which states that the average performance of prediction models becomes equivalent across all scenarios. The lack of a consistently superior model is likely a consequence of individual prediction models being well-suited for specific scenarios, but not for all. Hence, the continuous exploration for an all-rounder individual genomic prediction model may not be an optimal approach to improve overall prediction performance.

Combining multiple genomic prediction models into an ensemble can be an alternative approach to using each model independently (Cooper et al. 2025). The Diversity Prediction Theorem (Page 2018) indicates that the use of diverse prediction algorithms in a many-model ensemble lowers the prediction error more than the average prediction error of individual prediction models. In applications of genomic prediction for crop breeding, each genomic prediction model predicts phenotypes by capturing a different view of the genetic architecture based on the properties of their unique algorithms. The Diversity Prediction Theorem suggests that the aggregation of such diverse views can capture a more complete model of trait genetic architecture, enabling improved prediction performance (Messina et al. 2025; Tomura et al. 2025b).

We investigated whether an ensemble of diverse genomic prediction models improves prediction performance through the application to traits from two well-documented Nested Association Mapping (NAM) examples (Buckler et al. 2009; Chen et al. 2019). The parents of these two NAM datasets represent a wide range of maize genetic diversity, extending from domestication to modern inbred lines used for hybrid breeding (Hufford et al. 2012). This study was oriented by two objectives. Firstly, the prediction performance of the ensemble was compared with individual genomic prediction models to evaluate the expectations of the Diversity Prediction Theorem for prediction performance improvement. Secondly, the estimates of genomic marker effects from each genomic prediction model were compared to observe the association between the diversity in quantified marker effects and the prediction performance of the ensemble.

Materials and Methods

1. Datasets

The TeoNAM dataset (Chen et al. 2019) is a collection of samples from the five recombinant inbred line (RIL) populations between the maize line W22 and five teosinte types (TIL01, TIL03, TIL11 and TIL14 from *Z. mays ssp. parviglumis* and TIL25 from *Z.mays ssp. mexicana*). In each RIL population, the F1 was backcrossed with W22. After the backcross was completed, each population was self-pollinated four times to develop the RIL populations. A randomised complete block design was applied to test the RILs for each population contributing to the TeoNAM experiment. Each population was tested twice at the University of Wisconsin West Madison Agricultural Research Station. For W22TIL01, W22TIL03 and W22TIL11, the RILs were tested in the summers of 2015 and 2016. For W22TIL14, the RILs were tested in the summers of 2016 and 2017, and for W22TIL25 the RILs were tested in two different blocks in the summer of 2017.

The MaizeNAM dataset (Buckler et al. 2009) consists of samples from 25 RIL populations derived from crosses between the maize line B73 and a diverse set of 25 inbred lines from temperate and tropical regions. The 25 RIL populations were developed through self-pollination to the F5 generation. Each population was tested in the following locations during the summer in the United States: Aurora in New York, Clayton in North Carolina, Columbia in Missouri and Urbana in Illinois. Two experiments were conducted at each location with a randomised design and hence each population was evaluated in the eight environments. The recorded phenotype scores in each environment were leveraged to estimate the best linear unbiased predictions (BLUPs) with ASReml (v2.0) in each population. Those BLUPs were used as trait phenotypes of each RIL. The total number of genomic markers (SNPs) and RIL samples from each target dataset is summarised in Table 1.

The selection of the two maize NAM datasets was motivated by their different levels of genetic diversity. The TeoNAM dataset contains a high level of genetic diversity: Crosses of a temperate maize inbred with representatives of its ancestral species, Teosinte, uncover components of trait genetic architecture fixed during domestication. In contrast, the MaizeNAM dataset is less diverse, as crosses are limited to within a set of domesticated maize lines. The nucleotide diversity (π) of Teosinte (*Z. mays ssp. parviglumis*) used in the

TeoNAM populations was higher at 0.0059 per base pair across 10-kb windows, compared to that of improved maize lines included as parental lines in several MaizeNAM populations at 0.0048 (Hufford et al. 2012). Using the two maize NAM datasets characterised by different levels of genetic diversity, we investigated the distinct characteristics of the genomic prediction models for these datasets to understand the predictive behaviour of the individual models and their ensemble.

2. Data preprocessing

Data preprocessing methods leveraged in Tomura et al. (2025b) were also applied in this study. For the TeoNAM dataset, missing SNP markers were imputed by assigning genomic marker values based on flanking markers. RIL samples with missing genomic markers for an entire chromosome and those without phenotype values were removed. For the MaizeNAM datasets, missing SNP markers were imputed with flanking markers and missing phenotypes were removed by Buckler et al. (2009). Hence, no further imputation was required. Prior to the ensemble-based analyses, the total number of SNPs was reduced further based on linkage disequilibrium (LD) criteria using PLINK (v1.9) (Chang et al. 2015); Any SNP with a squared correlation of 0.8 or above was removed using a window size of 30,000 and a step size of 5 in both datasets.

For the TeoNAM dataset, the DTA and ASI traits for each RIL were recorded twice as reported by Chen et al. (2019). For the ensemble analysis, two trait phenotype records of each RIL were concatenated in the TeoNAM dataset following Tomura et al. (2025b). For genomic prediction, the data were merged with a factor distinguishing the two different environments.

3. Genetic variance component calculation

Both additive and non-additive (epistatic) variance components were estimated per population in each dataset following Yadav et al. (2021, 2023). Each variance component was calculated by constructing the respective genomic relationship matrices. The additive genomic relationship matrix (G_A) was defined below (Yang et al. 2010):

$$G_A = \frac{WW'}{n} \tag{1}$$

where W is the incidence matrix corresponding to the additive marker effects and n is the number of individuals. The epistatic genomic relationship matrix (G_{AA}) was developed by the Hadamard product ($\odot$) of the additive genomic relationship matrix as defined below:

$$G_{AA} = \frac{G_A \odot G_A}{tr(G_A \odot G_A)/_n} \tag{2}$$

where $tr(\cdot)$ is the trace that sums the diagonal elements in the matrix. Using the G_A and G_{AA} matrices, extended genomic best linear unbiased prediction (extended-GBLUP) was modelled:

$$y^{n*1} = X\beta + Wa + Wt + \varepsilon \tag{3}$$

where y is the vector of predicted phenotypes, X is the incidence matrix assigned to fixed effects β, W is the incidence matrix for the random effects, $a \sim N(0, G_A, \sigma_A^2)$ is the vector of additive genomic marker effects, $t \sim N(0, G_{AA}, \sigma_E^2)$ is the vector of epistatic genomic marker effects, and $\varepsilon \sim N(0, I, \sigma_\varepsilon^2)$ is the vector of random residual effects. The variance of the additive genomic marker effects (σ_A^2), epistatic genomic marker effects (σ_E^2) and random residual effects (σ_ε^2) were leveraged in the normalised form as a variance component in this study. The extended-GBLUP was implemented by ASReml-R (v4.2) (Butler et al. 2017).

4. Genomic prediction models

Six individual genomic prediction models were employed in this study. Three classical genomic prediction models, ridge regression best linear unbiased prediction (rrBLUP; Meuwissen et al. 2001), BayesB (Meuwissen et al. 2001) and reproducing kernel Hilbert space regression (RKHS; Gianola and Van Kaam 2008), were applied in this study. rrBLUP and BayesB were implemented applying a linear mixed model (Pérez and de Los Campos 2014):

$$\boldsymbol{\eta} = 1\boldsymbol{\mu} + \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j + \sum_{l=1}^{L} \boldsymbol{u}_l \tag{4}$$

where $\boldsymbol{\eta} = \{\eta_1, \eta_1, \ldots, \eta_n\}$ is the vector of the predicted phenotypes for observed phenotypes $\boldsymbol{y} = \{y_1, y_1, \ldots, y_n\}$, $\boldsymbol{\mu}$ is the intercept, $\boldsymbol{X}_j$ is the design matrix based on the marker values, $\boldsymbol{\beta}_j$ is the coefficient matrix storing genomic marker effects and $\boldsymbol{u}_l = \{u_{l1}, u_{l2}, u_{ln}\}$ is the vector of random effects. For rrBLUP, $\boldsymbol{\beta}_j$ is defined as $\boldsymbol{\beta}_j \sim N(0, \boldsymbol{I}, \sigma^2_{\beta_j})$ where $\boldsymbol{I}$ is the identity matrix and σ^2 is the variance of genomic marker effects (Endelman 2011). For BayesB, some of the genomic marker effects can be shrunk to zero instead of having a Gaussian distribution. For RKHS, a non-parametric component is added to Equation (4):

$$\boldsymbol{\eta} = 1\boldsymbol{\mu} + \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j + \sum_{j=1}^{J} g(\boldsymbol{X}_j) + \sum_{l=1}^{L} \boldsymbol{u}_l \tag{5}$$

where $g(\boldsymbol{X}_j)$ are the genetic effects from non-additive genomic marker effects (Howard et al. 2014). $g(\boldsymbol{X}_j)$ is implemented following:

$$g(.) = \alpha_0 + \sum_{n=1}^{N} \alpha_n K(., x_{jn}) \tag{6}$$

where $\boldsymbol{X}j = \{x_{j1}, x_{j2}, \ldots x_{jn}\}$ is the set of genomic markers, α_0 is the intercept, α_n is the coefficient assigned to x_n and K is the reproducing kernel. rrBLUP, BayesB and RKHS were implemented using BGLR (Pérez and de Los Campos, 2014) in R (v1.1.4). The number of iterations was set as 12,000 and the burn-in was 2,000 for the three models in both datasets. Default values were leveraged for all other parameters.

Three machine learning approaches, random forest (RF; Breiman, 2001), support vector regression (SVR; Drucker et al. 1996) and multi-layer perceptron (MLP; Rosenblatt, 1958), were selected. RF is one of the ensemble models that consists of numerous decision trees. Each decision tree is developed using a fraction of the given dataset and returns prediction values by traversing the tree with input attribute values. In regression, the average of predicted values from all the decision trees is leveraged as the final prediction value. SVR predicts values by fitting a hyperplane under the constraints following the equation:

$$U\left(\sum_{i=1}^{N} \xi_i^* + \sum_{i=1}^{N} \xi_i\right) + \frac{1}{2}(\boldsymbol{w}^t \boldsymbol{w})$$

$$s.t.\ \ y_i - (\boldsymbol{w}^t \boldsymbol{v}_i) - b \leq \epsilon + \xi_i \tag{7}$$

$$(\boldsymbol{w}^t \boldsymbol{v}_i) + b - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i^* \leq 0$$

$$\xi_i \leq 0$$

where U is the objective function, b is the intercept, $\boldsymbol{w}$ is the coefficient vector, ϵ determines the distance between a hyperplane and the decision boundary, $\boldsymbol{v}$ is the vector of data points and ξ_i^* and ξ_i are slack variables. The location of the hyperplane is optimised by maximising the number of samples included in the space between the hyperplane and the decision boundary. MLP mimics human brain systems by continuous nonlinear aggregation of information from neighbouring neurons in a forward direction (Dave and Kamlesh, 2014) such that:

$$z_i = f\left(\sum_j^J v_{ij} y_j + \theta_i\right) \tag{8}$$

where z_i is the output value from the neuron i, v_{ij} is the edge weight connecting the neuron i and neuron j in the previous layer, J represents the neurons in the previous layer, y_j is the predicted value from the neuron j in the previous layer, θ_i is the bias in the neuron i and $f(\cdot)$ is the activation function that nonlinearly converts the aggregated value. RF and SVR were implemented using Sklearn (v1.2.2) in Python (v3.11.10). The default hyperparameter values were leveraged for both prediction models in both datasets except for the number of trees (1,000) in RF. MLP was executed using pytorch (v1.13.1). In both datasets, MLP was structured with one hidden layer. For the TeoNAM dataset, 50 neurons were included in the hidden layer with a dropout of 0 and activated by the Rectified Linear Unit (ReLU; Nair et al. 2010) activation. The epoch of 200 was leveraged to train the model with a learning rate of 0.005 using the Adaptive Moment Estimation with Weight Decay (AdamW; Loshchilov and Hutter 2017) optimiser. For the MaizeNAM dataset, the number of neurons was reduced to 10 with a dropout of 0.1 and activated by the Rectified Linear Unit (ReLU) activation. The epoch 2,500 was used with the learning rate of 0.005 using RMSprop.

5. The Diversity Prediction Theorem

The ensemble-average model averaged the predicted phenotypes from the six individual models with equal weights using Python. The Diversity Prediction Theorem compares the error of the ensemble-average model against the average error of individual prediction models in relation to the diversity of the predicted values (Page 2018):

$$(\bar{M} - V)^2 = \sum_{i=1}^{N} \frac{(M_i - V)^2}{N} - \sum_{i=1}^{N} \frac{(M_i - \bar{M})^2}{N} \tag{9}$$

where M_i is the set of predicted values from the prediction model i, $\bar{M}$ is the set of mean predicted values from the i individual prediction models, V is the set of true values and N is the total number of prediction models considered. The many-model error (the first term) is calculated by subtracting the prediction diversity (third term) from the average error (second term). In our study, observed phenotypes are defined as the true value $V$and the many-model error is regarded as the ensemble error. $\bar{M}$ is calculated by averaging the predicted phenotypes from all the individual genomic prediction models used in this study.

Since the scale of the phenotypes varied across the traits and datasets, the value of the first term is not directly comparable. Hence, the ratio of the third term over the second term was leveraged as a metric to measure the diversity level of predicted phenotypes for the ensemble in each prediction scenario using Equation (9):

$$ratio = \frac{\sum_{i=1}^{N} \frac{(M_i - \bar{M})^2}{N}}{\sum_{i=1}^{N} \frac{(M_i - V)^2}{N}} \tag{10}$$

The ratio eliminates the scale difference between the phenotypes across prediction scenarios and hence the values of Equation (10) can be compared across the prediction scenarios.

6. Experimental design

The individual and ensemble genomic prediction models were evaluated by iterative prediction scenarios with different settings. For the TeoNAM dataset, three different training-test set ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5) were leveraged to mitigate the data size effect. In each population-training-test-set scenario, RIL data

were randomly sampled 500 times. Hence, 7,500 prediction scenarios (5 populations * 3 ratios * 500 samples) were developed per trait. The same three training-test set ratios were used in the MaizeNAM dataset. For each of the 25 populations in the dataset, the RILs were sampled 50 times for each population-training-test-set scenario, resulting in 3,750 prediction scenarios (25 populations * 3 ratios * 50 samples) per trait. In each prediction scenario, Pearson correlation, MSE and the three components of the Diversity Prediction Theorem (Equation (9)) were calculated. For the first two metrics, median values were calculated to develop violin plots, whereas the mean value was calculated for the three components of the Diversity Prediction Theorem.

7. Genomic marker effect estimation

To investigate the genetic architecture of each trait, we estimated and compared the genomic marker effects between the genomic prediction models. For rrBLUP and BayesB, $\boldsymbol{\beta}$ (representing the allele substitution effect) was extracted as the genomic marker effects. For RF, genomic marker effects were extracted using feature importance based on the impurity-based approach (Ishwaran 2015):

$$\hat{\Delta}(s,t) = \hat{\Delta}(t) - \left[\frac{N_L}{N}\hat{\Delta}(t_L) + \frac{N_R}{N}\hat{\Delta}(t_R)\right] \tag{11}$$

where s is a value from the attribute X, $\hat{\Delta}(t)$ is the impurity value at the node t, N is the total number of samples in the daughter node t_L and t_R, $\frac{N_L}{N}$ is the weight of the left daughter node t_L, $\frac{N_R}{N}$ is the weight of the right daughter node t_R , $\hat{\Delta}(t_L)$ is the impurity value of the left daughter node t_L and $\hat{\Delta}(t_R)$ is the impurity value of the right daughter node t_R. The importance of an attribute is determined by the total impurity value derived from the impurity value of each node for the attribute. Sklearn was employed to calculate the impurity values in Python.

For RKHS, SVR, MLP and genomic marker interaction in RF, Shapley values (Shapley 1953) were extracted as genomic marker effects (Lundberg and Lee 2017):

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_s(x_s)\right] \tag{12}$$

where Φ_i is the Shapley value for an attribute (SNP) i, S is a subset of features F, $f_{S\cup\{i\}}$ is a model prediction that predicts the target value (phenotype) with a feature subset including i, f_s is a model predicting phenotype excluding i. Shapley values estimate the causal effect of a particular SNP in the form of a conditional effect. For RKHS, the Shapley value was implemented using iml (v0.11.3) (Molnar et al. 2020) in R. SHAP (v0.42.1) (Lundberg and Lee 2017) was applied to SVR and MLP marker effects and RF marker-by-marker interactions in Python. Since Shapley values for each SNP were calculated per data point (RIL), Shapley values were summed across the 50 randomly chosen RILs in the test set to represent the genomic marker effects in the absolute value form in each prediction scenario.

8. Comparison with prior knowledge

We extracted information on genome positions of known genes and regulators for DTA from three previous studies. QTL regions were discovered by Chen et al. (2019) for the TeoNAM dataset and Buckler et al. (2009) for the MaizeNAM dataset. Wisser et al. (2019) identified several key genes affecting flowering time in maize in an empirical experiment of the short-term evolution of tropical region maize lines. Lastly, Dong et al. (2012) created a list of key gene regions and their interactions for flowering time in maize. They classified key gene regions into two categories (leaf and shoot apical meristem (SAM)) based on the tissue or organ of the maize plant where they act to determine flowering time. These genes were also classified into seven categories based on their a priori defined functions (light transduction, clock, photoperiod, autonomous, aging, Gibberellin (GA) and integrator). The comparison of known gene regulators and the estimated genomic marker effects with interactions was visualised by developing circos plots (Krzywinski et al. 2009; Tomura et al. 2025a) using pyCirclize (v1.7.1) in Python.

Results

1. Two maize NAM datasets and traits showed a distinctive composition of genetic variance components

The estimated genetic variance components for the traits provide a preliminary characterisation of the two target datasets (Fig. 1). The average ratio of the non-additive genetic variance component over the additive

genetic variance component for the TeoNAM dataset (0.450) was higher than for the MaizeNAM dataset (0.300) for DTA, indicating the potentially higher non-additive effects on DTA for the TeoNAM dataset. However, the level of the non-additive variance component ratio of DTA remained relatively small compared to ASI in both TeoNAM (0.610) and MaizeNAM dataset (0.626). It is more likely that ASI is influenced by higher non-additive genomic marker effects compared to DTA. Such a difference in the variance components can reveal distinctive predictive behaviours of the genomic prediction models in respective prediction scenarios, investigated in the following subsections.

2. Ensemble prediction outperformed the average of individual genomic prediction models

The ensemble-average model improved the prediction performance across the target traits and datasets by increasing accuracy and reducing error (Fig. 2). For the TeoNAM dataset, the ensemble-average model reached higher median prediction accuracy for DTA (0.842) and ASI (0.505) in contrast to the average prediction accuracy of the individual genomic prediction models for DTA (0.741) and ASI (0.473). The ensemble-average model also had a smaller median prediction error for DTA (the ensemble-average model = 11.033 and the average of individual genomic prediction models = 14.699) and ASI (the ensemble-average model = 4.339 and the average of individual genomic prediction models = 4.638). For the MaizeNAM dataset, the median prediction accuracy for DTA and ASI was higher for the ensemble-average model (0.642 and 0.465, respectively) compared to the average performance of the individual genomic prediction models (0.601 and 0.436, respectively). Similarly, there was improvement in the median prediction error for both traits (DTA = 2.389 and ASI = 0.442 in contrast to DTA = 2.978 and ASI = 0.464). The observed result indicates that the ensemble-average model consistently reached a higher prediction performance compared to the average performance of the individual genomic prediction models.

3. Diversity contributed to the performance improvement of the ensemble

The level of prediction improvement of the ensemble varied with traits and datasets, associated with the level of diversity in predicted phenotypes (Table 2). The ratio defined in Equation (10) based on the Diversity

298 Prediction Theorem was higher in the TeoNAM dataset for both traits (DTA = 0.42 and ASI = 0.19) in contrast
299 to the MaizeNAM dataset (DTA = 0.33 and ASI = 0.11) (Table 2). A higher ratio value implies that the
300 predicted phenotypes from the individual genomic prediction models are more diverse, relative to the basis of
301 the mean predicted phenotypes. Since greater improvements in prediction performance were observed in the
302 TeoNAM dataset compared to the MaizeNAM dataset, the higher ratio values for both DTA and ASI in the
303 TeoNAM dataset were associated with higher diversity in the predicted phenotypes from the individual
304 genomic prediction models. Additionally, the higher ratio for DTA compared to ASI, for both datasets,
305 indicates the potential for greater prediction performance improvement for DTA than for ASI in both datasets.
306 These findings indicate that there is a positive association between the level of diversity in the predicted
307 phenotypes and the prediction performance of the ensemble.

308 The comparison of the predicted DTA at both the phenotype and genotype levels also indicated a
309 positive correlation between information diversity and prediction performance (Fig. 3). Most of the pairwise
310 comparisons in the MaizeNAM dataset demonstrated stronger positive associations (average for predicted
311 phenotypes = 0.911, and average for predicted genomic marker effects = 0.578) (Table S1) compared to the
312 TeoNAM dataset (average for predicted phenotypes = 0.578, and average for predicted genomic marker effects
313 = 0.307). These results indicate that, for the TeoNAM dataset, the individual genomic prediction models
314 returned more diverse predicted phenotypes by assigning different genomic marker effects. This higher
315 diversity among the individual genomic prediction models for the TeoNAM dataset likely contributed to the
316 improved prediction performance of the ensemble model in comparison to the MaizeNAM dataset. The
317 relatively higher diversity in the TeoNAM dataset was also observed for ASI at both the phenotype and genome
318 levels (Fig. S1; Table S2). The MaizeNAM showed a higher positive association in both levels (average for
319 predicted phenotypes = 0.942, and average for predicted genomic marker effects = 0.671) compared to the
320 TeoNAM dataset (average for predicted phenotypes = 0.818, and average for predicted genomic marker effects
321 = 0.498). The difference in the level of positive association between the two datasets for ASI was smaller
322 compared to that for DTA, aligning with the smaller difference in the level of prediction performance
323 improvement by the ensemble between the two datasets for ASI. Such a small level of difference also
324 corresponded to the subtle difference in the ratio of the estimated variance components between the datasets

325 (Fig. 1). Hence, quantified genomic marker effects with high diversity may lead to improved prediction
326 performance of the ensemble.

327

328 4. Several key gene regulators were captured by the ensemble with the diverse trait genetic architecture

329 The constructed circos plots showed several overlaps between the genomic marker regions highlighted by the
330 genomic prediction models and known key gene regulators for the traits (Fig. 4). For DTA, the features of the
331 genomic prediction models were associated with several key maize flowering genes in both datasets. For
332 example, in the TeoNAM dataset, the region between 35 cM and 50 cM in chromosome 10 was repeatedly
333 targeted across genomic prediction models (with MLP as an exception). This region contains ZmCCT10,
334 involved in regulating flowering time in maize, known to upregulate the circadian clock as part of the
335 photoperiod pathway (Dong et al. 2012; Chen et al. 2019; Wisser et al. 2019). The region containing
336 ZmCCT10 also showed strong interactions with the region in chromosome 8 (between 60 cM and 80 cM).
337 This region includes ZCN8, a well-established gene controlling flowering in maize that is known to interact
338 with ZmCCT10 (Dong et al. 2012; Chen et al. 2019; Wisser et al. 2019). In the MaizeNAM dataset, in addition
339 to the same two regions identified in the TeoNAM dataset, the genomic prediction models identified another
340 region between 50 cM and 60 cM in chromosome 9 that strongly interacted with the region containing
341 ZmCCT10 in chromosome 10 and the region containing ZCN8 in chromosome 8. This chromosome 9 region
342 contains ZmCCT9 (Wisser et al. 2019), another gene involved in the photoperiod pathway that negatively
343 regulates ZCN8 (Huang et al. 2018). For ASI, several overlaps between the highlighted gene regions by the
344 genomic prediction models and QTL regions (Buckler et al. 2009; Chen et al. 2019) were also observed in
345 both datasets (Fig. S2). Such an overlap between the gene regions identified across genomic prediction models
346 and the key gene regulators from the literature in both datasets indicates that these genomic prediction models
347 reliably identified regions containing several known causative loci.

348 The comparative analysis of the extracted genomic prediction model features revealed the different
349 components of trait genetic architecture captured by the alternative genomic prediction models. The pattern
350 of the genetic architecture from each individual genomic prediction model was more variable in the TeoNAM

351 dataset (Fig. 4a, S2a) in comparison to the MaizeNAM dataset (Fig. 4b, S2b), consistent with the Diversity
352 Prediction Theorem framework (Table 2, Fig. 3, S1) and the results of the ensemble model prediction. While
353 the different genomic prediction models identified many similar regions of the genome for DTA, the
354 contributions of each marker identified in these regions could vary considerably in the representation of the
355 genetic architecture in the TeoNAM dataset (Fig. 4a, S2a). In the MaizeNAM dataset, on the other hand, there
356 was more consensus across the genomic prediction models (Fig. 4b, S2b). The genomic prediction models
357 detected similar genomic regions with a stronger consensus of the contribution of each marker. It is also noted
358 that, for the MaizeNAM dataset, the genomic prediction models allocated stronger weights to the markers at
359 the peaks of the QTL identified by Buckler et al. (2009) from their analysis of the same dataset (Fig. 3b; lower
360 triangle of pairwise comparisons). In contrast, for the TeoNAM dataset, while the markers at the peaks of the
361 QTL identified by Chen et al. (2019) were included in the genomic prediction models, stronger weights were
362 less frequently assigned by the genomic prediction models (Fig. 3a; lower triangle of pairwise comparisons).
363 The analysis of the genomic prediction models through the view of the trait genetic architecture visualised the
364 different diversity levels in the inferred genetic architecture.

365

366 Discussion

367 Breeders have continuously searched for prediction methods that can consistently and accurately distinguish
368 individuals with the desired traits in their breeding programs (Voss-Fels et al. 2019; Crossa et al. 2025;
369 Escamilla et al. 2025). Genomic prediction models can accurately identify these individuals under certain
370 prediction scenarios (Azodi et al. 2019). However, a priori, it is often not possible to obtain information that
371 determines the optimal combination of scenarios and genomic prediction models without a significant trial-
372 and-error process (Jeon et al. 2023; Crossa et al. 2025). In contrast, the ensemble of multiple individual
373 genomic prediction models consistently improved prediction performance in accordance with the expectations
374 of the Diversity Prediction Theorem. The improvement in prediction performance achieved by the ensemble
375 was marginal in some cases, depending on the diversity in the predictive information included in the individual
376 prediction models. However, the performance of the ensemble approach was always at least comparable to
377 the average prediction performance of individual genomic prediction models. This observation aligned with

the conclusions from other empirical investigations of ensemble studies for genomic prediction in crop breeding (Wallach et al. 2018; Heilmann et al. 2023; Kick and Washburn 2023; Tomura et al. 2025b). The prediction performance of the ensemble models in their studies reached the highest or equivalent to that of the best individual prediction model. Hence, both theory and the body of empirical evidence indicate that an ensemble of genomic prediction models may be a more efficient approach than continuous trial-and-error investigation to determine the best individual genomic prediction model for each trait and dataset.

The success of the ensemble approach demonstrated in this study is interpreted in part a consequence of the high levels of genetic diversity represented in both datasets, especially in the TeoNAM (Buckler et al. 2009; Hufford et al. 2012; Chen et al. 2019). The high levels of genetic variation in both NAM experiments contributed to the high levels of diversity in the predicted phenotypes and the estimated genomic marker effects from the individual genomic prediction models. While high diversity in predicted values can improve the performance of the ensemble approach, excessively diverse prediction values may not always lead to further increases in the prediction performance. The level of diversity needs to be constrained within a certain range that varies by prediction scenarios (Bian and Chen 2021; Wood et al. 2023). The benefits of such constrained levels of diversity align with the implications of the theoretical framework of the Diversity Prediction Theorem (Page 2018). This framework can be applied to investigate the influence of diverse predicted values (Equation (9) third term) towards the prediction error of the ensemble (Equation (9) first term) in relation to the average of predicted values (Equation (9) second term) from each prediction model, as was conducted for the TeoNAM and MaizeNAM datasets here.

Given that prediction performance improvement was observed for the simplest ensemble approach (averaging predicted phenotypes with equal weights), there is potential for further performance improvement using advanced ensemble design algorithms. One promising approach can be weight optimisation. The minimisation of the Many-Model error in the Diversity Prediction Theorem (Page 2018) can be leveraged as one method to optimise the weights. This and other forms of ensemble optimisation are an active area for further research. The leveraging of prior knowledge into the ensembles is another approach to improve the prediction performance of the ensembles. Our investigation of the TeoNAM and MaizeNAM datasets captured features of several known key genes (ZmCCT9, ZmCCT10 and ZCN8) and emphasised the importance of

405 other previously unknown genomic regions that are novel regions to discover new causative loci. Such
406 extracted models of trait genetic architecture can be leveraged as prior knowledge to supervise the training of
407 genomic prediction models. Prior knowledge can provide prediction models with external information to
408 integrate natural biophysical laws within their prediction mechanisms (Technow et al. 2015; Messina et al.
409 2018, 2025; Von Rueden et al. 2021; Cooper et al. 2025). Such applications of prior knowledge can mitigate
410 the risk of overfitting prediction models, especially when the training set is small and the prediction models
411 cannot be well-trained only using the training set (Schapire et al. 2002). Hence, the prediction performance of
412 individual genomic prediction models is expected to increase and, consequently, improve the prediction
413 performance of ensembles of these models.

414 In addition to revealing key features of the genetic architecture for traits measured in the TeoNAM and
415 MaizeNAM datasets, the ensemble of diverse genomic prediction models is demonstrated as a powerful
416 approach for more consistently achieving high prediction performance across breeding scenarios. Such
417 improvements in prediction performance enable the consistent identification of superior individuals with the
418 desired trait combinations for all stages of breeding programs. Future studies will explore opportunities to
419 extend the ensemble-average model used in this study, including the assessment of prediction performance
420 improvement by optimising weightings of models and integrating prior knowledge (Messina et al. 2025).
421 Overall, the ensemble of multiple genomic prediction models has the potential to improve breeding programs
422 by accelerating genetic gain.

423

424 Data and code availability

425 The TeoNAM dataset (Chen et al. 2019) used in this study is located at https://datacommons.cyverse.org/bro-
426 wse/iplant/home/shared/panzea/genotypes/GBS/TeosinteNAM for the genotypes and https://gsajournals.figs-
427 hare.com/articles/dataset/Supplemental_Material_for_Chen_et_al_2019/9250682 for the phenotypes. The
428 MaizeNAM dataset (Buckler et al. 2009) used in this study is located at https://cbsusrv04.tc.cornell.edu/use-
429 rs/panzea/download.aspx?filegroupid=10 for the genotypes and phenotypes. The code used in this study is

located at https://github.com/ShunichiroT/ensemble for the ensemble and https://github.com/ShunichiroT/Circos_plot for circos plots.

Author contribution

ST designed the experiment, analysed data, visualised the results and wrote the manuscript. OP and MJW provided feedback on the analysis, results and manuscript draft. MC formalised the concept, introduced the datasets, designed the experiment and provided feedback on the analysis, results and manuscript draft. All authors discussed the results.

Acknowledgments

We thank the National Computing Infrastructure (NCI) and the Research Computing Centre (RCC) at the University of Queensland for providing access to the High Performance Computing (HPC) machines. We also thank Dr Seema Yadav for the support in the variance component analysis.

Funding

This study was funded by the Australian Research Council through the support of the Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture (CE200100015).

Conflicts of interest

The authors declare no conflicts of interest.

References

Alemu A, Åstrand J, Montesinos-Lopez OA, y Sanchez JI, Fernandez-Gonzalez J, Tadesse W, Vetukuri RR, Carlsson AS, Ceplitis A, Crossa J *et al*. (2024) Genomic selection in plant breeding: Key factors shaping two decades of progress. Molecular Plant. 17:552–578. 10.1016/j.molp.2024.03.007.

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics. 9:3691–3702. https://doi.org/10.1534/g3.119.400498.

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Science. 47:1082–1090. https://doi.org/10.2135/cropsci2006.11.0690.

Bian Y, Chen H (2021) When does diversity help generalization in classification ensembles? IEEE Transactions on Cybernetics. 52:9059–9075. https://doi.org/10.1109/TCYB.2021.3053165.

Breiman L. (2001) Random forests. Machine Learning. 45:5–32. https://doi.org/10.1023/A:1010933404324.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al*. (2009) The genetic architecture of maize flowering time. Science. 325:714–718.

Butler D, Cullis B, Gilmour A, Gogel B, Thompson R (2017) Asreml-r reference manual version 4. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Ceccarelli S, Grando S, Maatougui M, Michael M, Slash M, Haghparast R, Rahmanian M, Taheri A, Al-Yassin A, Benbelkacem A *et al*. (2010) Plant breeding and climate changes. The Journal of Agricultural Science. 148:627–637. https://doi.org/10.1017/S0021859610000651.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation plink: rising to the challenge of larger and richer datasets. GigaScience. 4. https://doi.org/10.1186/s13742015-0047-8.

Chen Q, Yang CJ, York AM, Xue W, Daskalska LL, DeValk CA, Krueger KW, Lawton SB, Spiegelberg BG, Schnell JM *et al*. (2019) Teonam: A nested association mapping population for domestication and agronomic trait analysis in maize. Genetics. 213:1065–1078. https://doi.org/10.1534/genetics.119.302594.

Cooper M, Messina CD (2023) Breeding crops for drought-affected environments and improved climate resilience. The Plant Cell. 35:162–186. https://doi.org/10.1093/plcell/koac321.

Cooper M, Tomura S, Wilkinson MJ, Powell O, Messina CD (2025) Breeding perspectives on tackling trait genome-to-phenome (g2p) dimensionality using ensemble-based genomic prediction. Theoretical and Applied Genetics. https://doi.org/10.1007/s00122-025-04960-6.

Crossa J, Montesinos-Lopez OA, Costa-Neto G, Vitale P, Martini JW, Runcie D, Fritsche-Neto R, Montesinos-Lopez A, Pérez-Rodríguez P, Gerard G *et al*. (2025) Machine learning algorithms translate big data into predictive breeding accuracy. Trends in Plant Science. https://doi.org/10.1016/j.tplants.2024.09.011.

Dave VS, Dutta K (2014) Neural network based models for software effort estimation: a review. Artificial Intelligence Review. 42:295–307. https://doi.org/10.1007/s10462-012-9339-x.

Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M (2012) A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. Plos One. https://doi.org/10.1371/journal.pone.0043450.

Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V (1996) Support vector regression machines. Advances in neural information processing systems. 9.

Endelman JB (2011) Ridge regression and other kernels for genomic selection with r package rrblup. The Plant Genome. 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024.

Escamilla DM, Li D, Negus KL, Kappelmann KL, Kusmec A, Vanous AE, Schnable PS, Li X, Yu J (2025) Genomic selection: Essence, applications, and prospects. The Plant Genome 18:e70053. https://doi.org/10.1002/tpg2.70053

Gianola D, Van Kaam JB (2008) Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 178:2289–2303. https://doi.org/10.1534/genetics.107.084285.

Heffner EL, Lorenz AJ, Jannink JL, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop science. 50:1681–1690. https://doi.org/10.2135/cropsci2009.11.0662.

Heilmann PG, Frisch M, Abbadi A, Kox T, Herzog E (2023) Stacked ensembles on basis of parentage information can predict hybrid performance with an accuracy comparable to marker-based gblup. Frontiers in Plant Science. 14:1178902.

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop science. 52:146–160. https://doi.org/10.2135/cropsci2011.06.0297.

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes—Genomes—Genetics. 4:1027–1046. https://doi.org/10.1534/g3.114.010298.

Huang C, Sun H, Xu D, Chen Q, Liang Y, Wang X, Xu G, Tian J, Wang C, Li D *et al*. (2018) Zmcct9 enhances maize adaptation to higher latitudes. Proceedings of the National Academy of Sciences. 115:E334–E341. https://doi.org/10.1073/pnas.1718058115.

Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM *et al*. (2012) Comparative population genomics of maize domestication and improvement. Nature genetics. 44:808–811. https://doi.org/10.1038/ng.2309.

Ishwaran H (2015) The effect of splitting on random forests. Machine learning. 99:75–118. https://doi.org/10.1007/s10994-014-5451-2.

Jeon D, Kang Y, Lee S, Choi S, Sung Y, Lee TH, Kim C (2023) Digitalizing breeding in plants: A new trend of next-generation breeding based on genomic prediction. Frontiers in Plant Science. 14:1092584. https://doi.org/10.3389/fpls.2023.1092584.

John M, Haselbeck F, Dass R, Malisi C, Ricca P, Dreischer C, Schultheiss SJ, Grimm DG (2022) A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. Frontiers in Plant Science. 13:932512. https://doi.org/10.3389/fpls.2022.932512.

Kholová J, Urban MO, Cock J, Arcos J, Arnaud E, Aytekin D, Azevedo V, Barnes AP, Ceccarelli S, Chavarriaga P *et al*. (2021) In pursuit of a better world: crop improvement and the cgiar. Journal of Experimental Botany. 72:5158–5179. https://doi.org/10.1093/jxb/erab226.

Kick DR, Washburn JD (2023) Ensemble of best linear unbiased predictor, machine learning, and deep learning models predict maize yield better than each model alone. in silico Plants. p. diad015. https://doi.org/10.1093/insilicoplants/diad015.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome research. 19:1639–1645. https://doi.org/10.1126/science.1174276.

Langridge P, Braun H, Hulke B, Ober E, Prasanna B (2021) Breeding crops for climate resilience. Theoretical and Applied Genetics. 134:1607–1611. https://doi.org/10.1007/s00122-021-03854-7.

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems. 30.

Messina C, Garcia-Abadillo J, Powell O, Tomura S, Zare A, Ganapathysubramanian B, Cooper M (2025) Toward a general framework for ai-enabled prediction in crop improvement. Theoretical and Applied Genetics. 138:1–15. https://doi.org/10.1007/s00122-025-04928-6.

Messina CD, Gho C, Hammer GL, Tang T, Cooper M (2023) Two decades of harnessing standing genetic variation for physiological traits to improve drought tolerance in maize. Journal of experimental botany. 74:4847–4861. https://doi.org/10.1093/jxb/erad231.

Messina CD, Technow F, Tang T, Totir R, Gho C, Cooper M (2018) Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (cgm) with whole genome prediction (wgp). European Journal of Agronomy. 100:151–162. https://doi.org/10.1016/j.eja.2018.01.007.

Meuwissen TH, Hayes BJ, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. genetics. 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Molnar C (2020) *Interpretable machine learning*. Lulu. com.

Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. pp. 807–814.

Page SE (2018) *The model thinker: What you need to know to make data work for you*. Hachette UK.

Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the bglr statistical package. Genetics. 198:483–495. https://doi.org/10.1534/genetics.114.164442.

Pixley KV, Cairns JE, Lopez-Ridaura S, Ojiewo CO, Dawud MA, Drabo I, Mindaye T, Nebie B, Asea G, Das B *et al*. (2023) Redesigning crop varieties to win the race between climate change and food security. Molecular Plant. 16:1590–1611. 10.1016/j.molp.2023.09.003.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 65:386. https://doi.org/10.1037/h0042519.

Schapire RE, Rochery M, Rahim M, Gupta N (2002) Incorporating prior knowledge into boosting. In: . volume 2. pp. 538–545.

Shapley LS (1953) Stochastic games. Proceedings of the national academy of sciences. 39:1095–1100. https://doi.org/10.1073/pnas.39.10.1095.

Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating crop growth models with whole genome prediction through approximate bayesian computation. PloS one. 10:e0130855. https://doi.org/10.1371/journal.pone.0130855.

Tomura S, Powell O, Wilkinson MJ, Cooper M (2025a) Circos plots for genome level interpretation of genomic prediction models. BioRxiv. https://doi.org/10.1101/2025.05.25.656055.

Tomura S, Wilkinson MJ, Cooper M, Powell O (2025b) Improved genomic prediction performance with ensembles of diverse models. G3: Genes, Genomes, Genetics. p. jkaf048. https://doi.org/10.1093/g3journal/jkaf048.

Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Pfrommer J, Pick A, Ramamurthy R *et al*. (2021) Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering. 35:614–633. https://doi.org/10.1109/TKDE.2021.3079836.

Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics. 132:669–686. https://doi.org/10.1007/s00122-018-3270-8.

Wallach D, Martre P, Liu B, Asseng S, Ewert F, Thorburn PJ, van Ittersum M, Aggarwal PK, Ahmed M, Basso B *et al*. (2018) Multimodel ensembles improve predictions of crop–environment–management interactions. Global Change Biology. 24. https://doi.org/10.1111/gcb.14411.

Werner CR, Zaman-Allah M, Assefa T, Cairns JE, Atlin GN (2025) Accelerating genetic gain through early-stage on-farm sparse testing. Trends in Plant Science. 10.1016/j.tplants.2024.10.010.

Wisser RJ, Fang Z, Holland JB, Teixeira JE, Dougherty J, Weldekidan T, de Leon N, Flint-Garcia S, Lauter N, Murray SC *et al*. (2019) The genomic basis for short-term evolution of environmental adaptation in maize. Genetics. 213:1479–1494. https://doi.org/10.1534/genetics.119.302780.

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE transactions on evolutionary computation. 1:67–82. https://doi.org/10.1109/4235.585893.

Wood D, Mu T, Webb AM, Reeve HW, Lujan M, Brown G (2023) A unified theory of diversity in ensemble learning. Journal of machine learning research. 24:1–49.

Yadav S, Ross EM, Wei X, Powell O, Hivert V, Hickey LT, Atkin F, Deomano E, Aitken KS, Voss-Fels KP *et al*. (2023) Optimising clonal performance in sugarcane: leveraging non-additive effects via mate-allocation strategies. Frontiers in Plant Science. 14:1260517. https://doi.org/10.3389/fpls.2023.1260517.

Yadav S, Wei X, Joyce P, Atkin F, Deomano E, Sun Y, Nguyen LT, Ross EM, Cavallaro T, Aitken KS *et al*. (2021) Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. Theoretical and Applied Genetics. 134:2235–2252. https://doi.org/10.1007/s00122-021-03822-1.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al*. (2010) Common snps explain a large proportion of the heritability for human height. Nature genetics. 42:565–569. https://doi.org/10.1038/ng.608.

601 Tables and Figures

602

603 **Table 1:** The total number of genomic markers (SNPs) before and after linkage disequilibrium (LD) filtering and the number
604 of recombinant inbred lines (RILs) for the TeoNAM and MaizeNAM datasets.

Dataset	SNPs		RILs
	Original	LD Filtering	
TeoNAM	11,396 - 16,110	268 - 342	219 - 308
MaizeNAM	1,106	170 - 251	126 - 196

605

606

607 **Table 2:** The mean value of each term in the Diversity Prediction Theorem for the days to anthesis (DTA) and anthesis to
608 silking interval (ASI) traits in the TeoNAM and MaizeNAM datasets. The ensemble error, average error and prediction
609 diversity represent the first, second and third term in the theorem, respectively. The third/second term is the ratio of the mean
610 prediction diversity over the mean average error.

611

Traits	Term	Term Description	TeoNAM	MaizeNAM
DTA	First	Ensemble error	14.36	3.38
	Second	Average error	24.93	5.05
	Third	Prediction diversity	10.57	1.67
	Third/Second	Ratio	0.42	0.33
ASI	First	Ensemble error	4.30	0.58
	Second	Average error	5.29	0.66
	Third	Prediction diversity	0.99	0.07
	Third/Second	Ratio	0.19	0.11

612

613

614

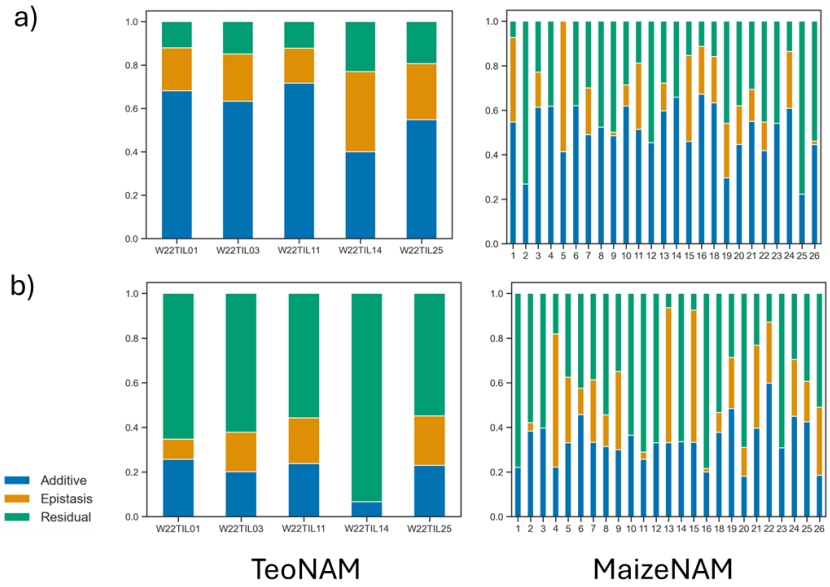

Fig. 1: The variance components of each population in the TeoNAM and MaizeNAM datasets for the a) days to anthesis (DTA) and b) anthesis to silking interval (ASI) trait. The bottom (blue), middle (orange) and top (green) parts on each bar represent the relative magnitudes of the additive, epistasis and residual variance components, respectively.

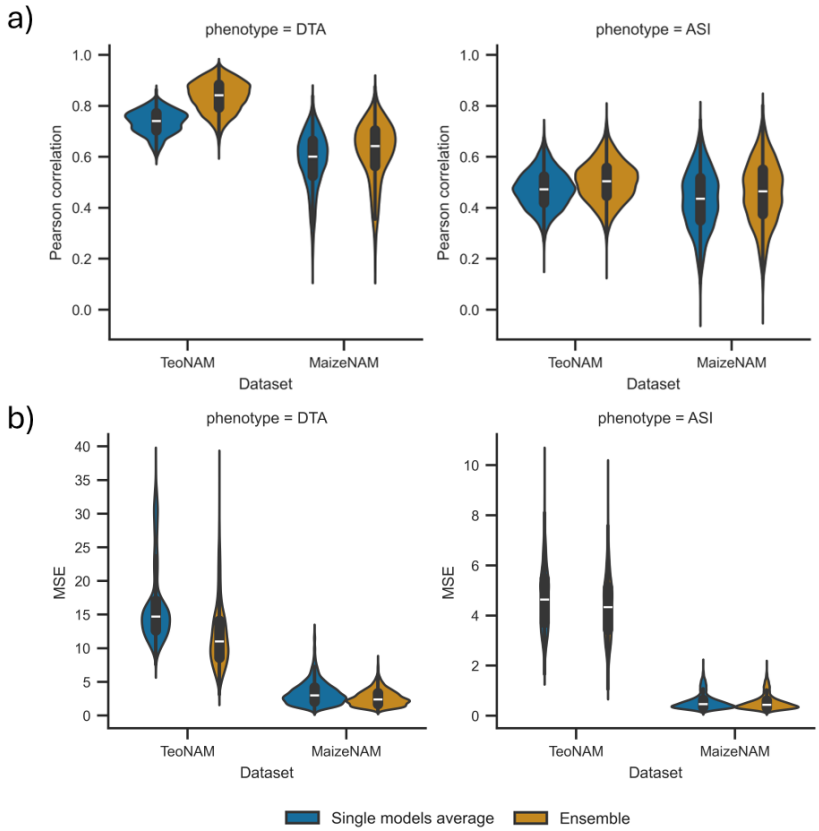

Fig. 2: The comparison of prediction performance between the average performance of the individual genomic prediction models (Single models average) and the ensemble-average model (Ensemble) for the days to anthesis (DTA) and anthesis to silking interval (ASI) traits in the TeoNAM and MaizeNAM datasets. The prediction performance was measured using the two metrics: a) Pearson correlation and b) mean square error (MSE). The prediction performance was measured in 7,500 prediction scenarios for the TeoNAM dataset and 3,750 prediction scenarios for the MaizeNAM dataset for each trait. The prediction scenarios were generated by the combination of the three training-test ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5), populations and sampling numbers. The width of the violins represents the distribution of performance metrics. The white horizontal lines on the black box plots show the median value for each metric. The whiskers extend 1.5 times the interquartile range.

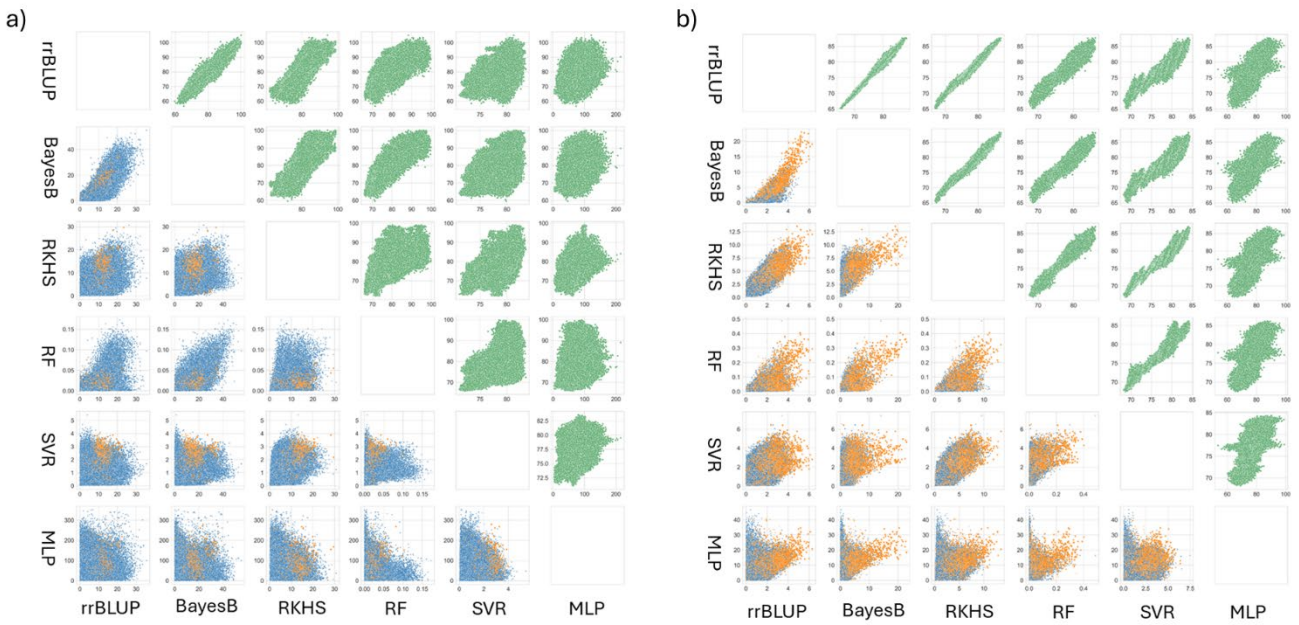

Fig. 3: Pairwise comparisons of the individual genomic prediction models for the days to anthesis (DTA) trait across all the prediction scenarios for the a) TeoNAM (7,500 prediction scenarios) and b) MaizeNAM (3,750 prediction scenarios) datasets. The prediction scenarios were generated by the combination of the three training-test ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5), populations and sampling numbers. The genomic prediction models were compared at the level of predicted phenotypes (top right triangle) and genomic marker effects (the bottom left triangle). The green dots represent a pair of predicted phenotypes of RIL samples in the test sets for each prediction scenario. The blue squares and orange triangles represent a pair of extracted genomic marker effects from each genomic marker in each sample scenario that were identified as non-QTL and QTL markers, respectively, by Chen et al. (2019) for the TeoNAM dataset and Buckler et al. (2009) for the MaizeNAM dataset.

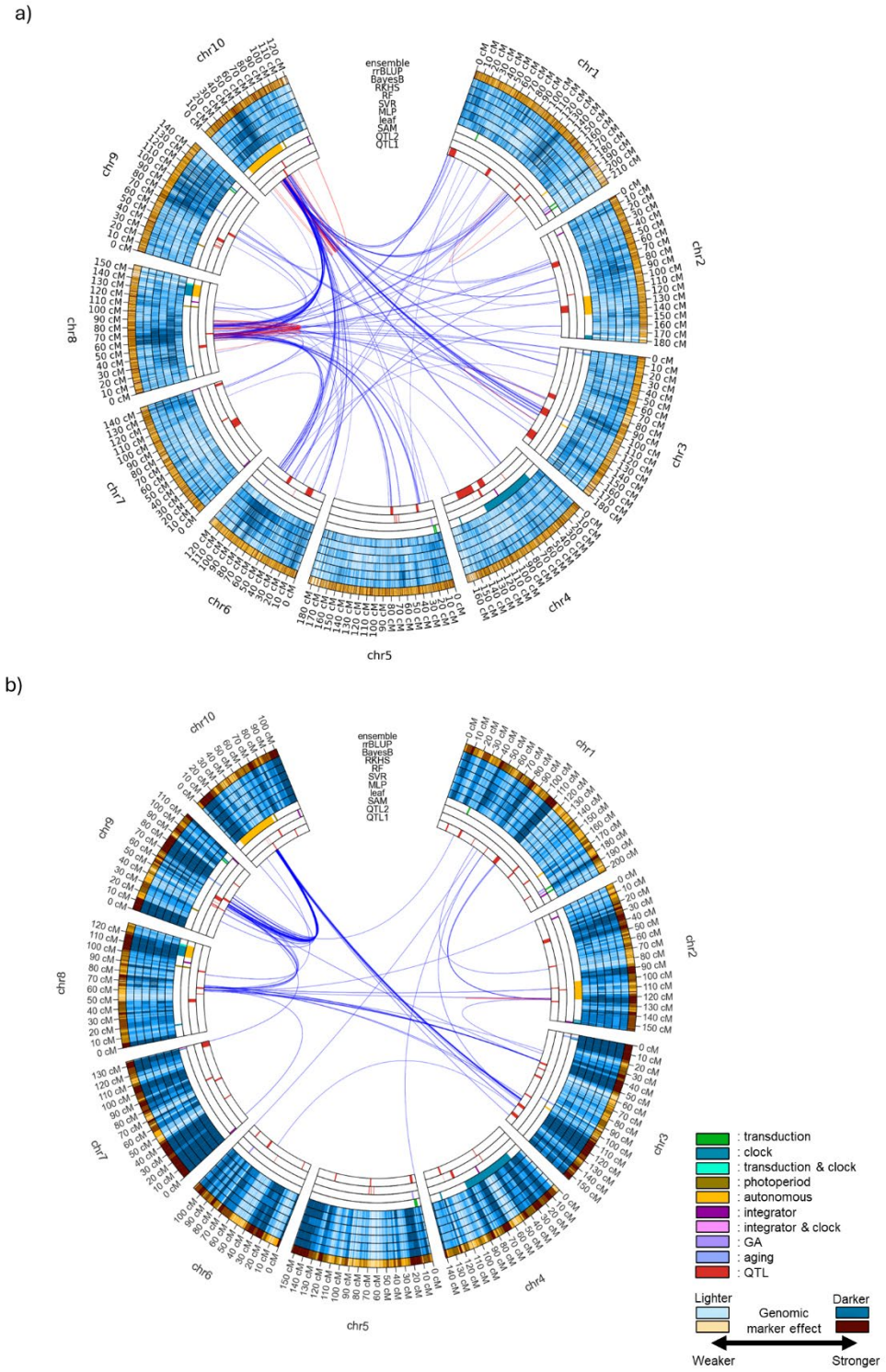

Fig. 4: Circos plots for the days to anthesis (DTA) trait for the a) TeoNAM and b) MaizeNAM datasets. The innermost (QTL 1) ring shows the QTL gene regions estimated by Chen et al. (2019) for the TeoNAM dataset and by Buckler et al. (2009) for the MaizeNAM dataset. The second innermost ring (QTL 2) represents the QTL gene regions identified by Wisser et al. (2019). The third and fourth innermost rings represent gene regulators that affect the shoot apical meristem (SAM) and leaf, respectively, identified by Dong et al. (2012). The blue rings (fifth to tenth) represent the genomic marker effects across the gene regions estimated by MLP, SVR, RF, RKHS, BayesB and rrBLUP, respectively. The outermost orange ring is the genomic marker effect for the ensemble-average model (ensemble). The darkness level of the blue and orange colours indicates the strength of the genomic marker effects, sectioned into ten levels using the quantiles. Darker colours represent higher genomic marker effect levels. The red and blue lines between genome regions are the genomic marker interaction effects calculated by pairwise Shapley scores from RF (top 0.01%; red = within chromosome and blue = between chromosomes).

Supplementary Materials

Table S1: Pearson correlation of each pair of individual genomic prediction models at the level of the predicted phenotypes (the top right triangle) and the genomic marker effect (the bottom left triangle) for the days to anthesis (DTA) trait in the a) TeoNAM and b) MaizeNAM datasets. Each value represents the correlation of its corresponding subplot in Figure 2.

a)

	rrBLUP	**BayesB**	**RKHS**	**RF**	**SVR**	**MLP**
rrBLUP		0.976	0.807	0.823	0.408	0.525
BayesB	0.738		0.762	0.884	0.381	0.512
RKHS	0.396	0.497		0.569	0.619	0.452
RF	0.355	0.629	0.406		0.310	0.408
SVR	0.143	0.258	0.559	0.332		0.231
MLP	0.062	0.062	0.071	0.036	0.062	

b)

	rrBLUP	**BayesB**	**RKHS**	**RF**	**SVR**	**MLP**
rrBLUP		0.997	0.994	0.968	0.960	0.803
BayesB	0.869		0.991	0.971	0.957	0.801
RKHS	0.886	0.786		0.981	0.981	0.787
RF	0.393	0.579	0.440		0.978	0.758
SVR	0.583	0.541	0.746	0.457		0.743
MLP	0.639	0.567	0.581	0.206	0.397	

Table S2: Pearson correlation of each individual genomic prediction model pair at the level of the predicted phenotypes (the top right triangle) and the genomic marker effect (the bottom left triangle) for the anthesis to silking interval (ASI) trait in the a) TeoNAM and b) MaizeNAM datasets. Each value represents the correlation of its corresponding subplot in Figure S2.

a)

	rrBLUP	**BayesB**	**RKHS**	**RF**	**SVR**	**MLP**
rrBLUP		0.991	0.942	0.839	0.850	0.661
BayesB	0.899		0.931	0.841	0.842	0.654
RKHS	0.777	0.716		0.927	0.896	0.715
RF	0.437	0.561	0.417		0.845	0.718
SVR	0.532	0.508	0.724	0.397		0.621
MLP	0.356	0.341	0.353	0.193	0.267	

b)

	rrBLUP	BayesB	RKHS	RF	SVR	MLP
rrBLUP		0.994	0.989	0.929	0.958	0.905
BayesB	0.505		0.985	0.941	0.956	0.900
RKHS	0.398	0.641		0.954	0.985	0.910
RF	0.478	0.708	0.712		0.961	0.867
SVR	0.623	0.667	0.662	0.903		0.892
MLP	0.507	0.822	0.719	0.894	0.828	

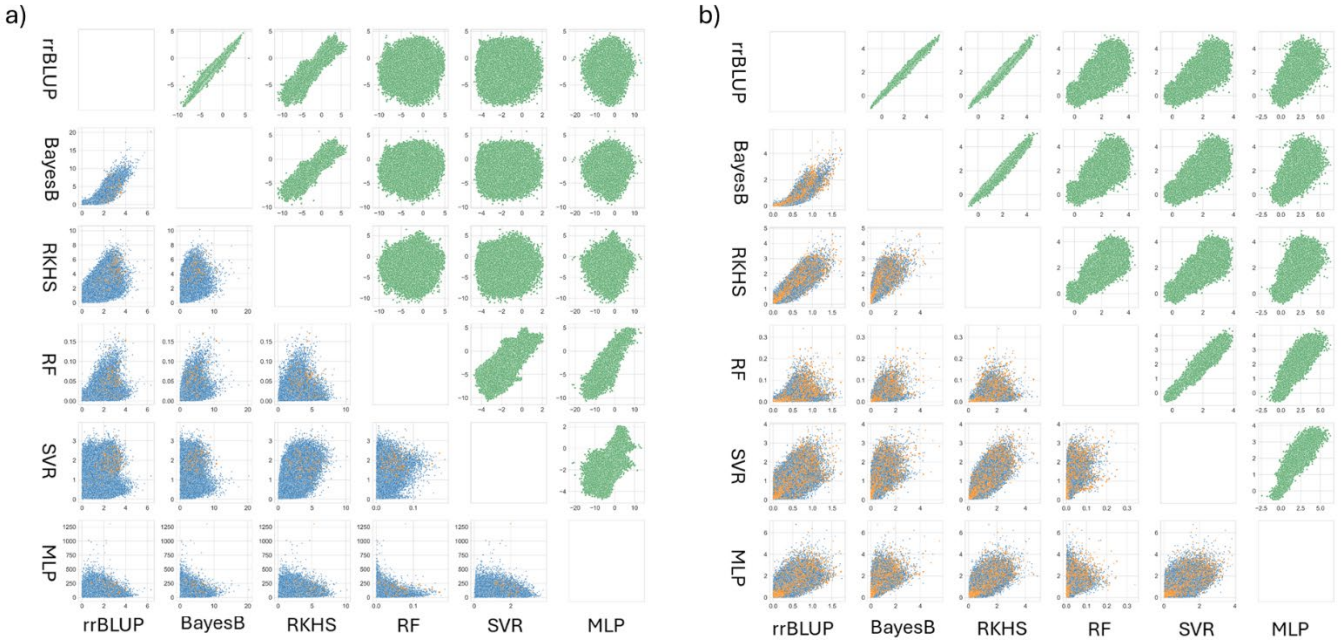

Fig. S1: Pairwise comparisons of the individual genomic prediction models for the anthesis to silking interval (ASI) trait across all the prediction scenarios for the a) TeoNAM (7,500 prediction scenarios) and b) MaizeNAM (3,750 prediction scenarios) datasets. The prediction scenarios were generated by the combination of the three training-test ratios (0.8-0.2, 0.65-0.35 and 0.5-0.5), populations and sampling numbers. The genomic prediction models were compared at the level of predicted phenotypes (top right triangle) and genomic marker effects (the bottom left triangle). The green dots represent a pair of predicted phenotypes of RIL samples in the test sets for each prediction scenario. The blue squares and orange triangles represent a pair of extracted genomic marker effects from each genomic marker in each sample scenario that were identified as non-QTL and QTL markers, respectively, by Chen et al. (2019) for the TeoNAM dataset and Buckler et al. (2009) for the MaizeNAM dataset.

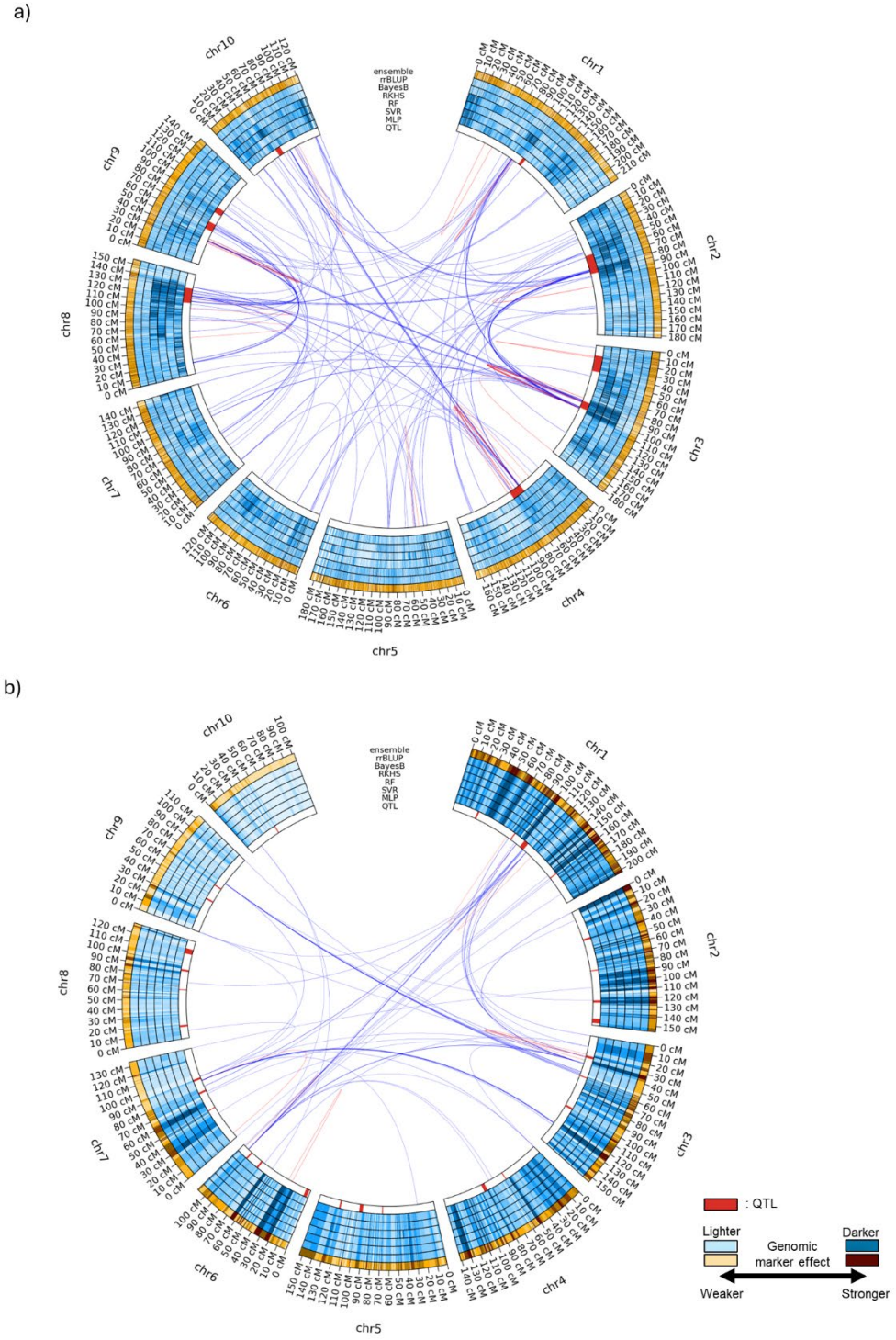

Fig. S2: Circos plots for the anthesis to silking interval (ASI) trait for the a) TeoNAM and b) MaizeNAM datasets. The innermost (QTL) ring shows the QTL gene regions estimated by Chen et al. (2019) for the TeoNAM dataset and by Buckler et al. (2009) for the MaizeNAM dataset. The blue rings (second to seventh) represent the genomic marker effects across the gene regions estimated by MLP, SVR, RF, RKHS, BayesB and rrBLUP, respectively. The outermost orange ring is the genomic marker effect for the ensemble-average model (ensemble). The darkness level of the blue and orange colours indicates the strength of the genomic marker effects, sectioned into ten levels using the quantiles. Darker colours represent higher genomic marker effect levels. The red and blue lines between genome regions are the genomic marker interaction effects calculated by pairwise Shapley scores from RF (top 0.01%; red = within chromosome and blue = between chromosomes).