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 result 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. While no individual prediction model outperforms all others, ensembles of multiple individual genomic prediction models can be applied to leverage their complementary strengths. The Diversity Prediction Theorem provides a framework for investigating the reduction in prediction error through the inclusion of diverse prediction models in the ensemble. We used the EasiGP (Ensemble AnalySis with Interpretable Genomic Prediction) software to investigate the performance of an ensemble approach, targeting 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 included as features in the different genomic prediction models with diverse weights among the individual models, demonstrating different views of the trait genetic architecture. The Diversity Prediction Theorem framework suggests that the ensemble combination of such diverse views likely 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, ensemble-based prediction can be applied to overcome some limitations proposed by the No Free Lunch Theorem when applying individual genomic prediction models.

Key words

Genomic prediction, ensemble, Diversity Prediction Theorem, No Free Lunch Theorem, maize, flowering time, genetic architecture

Article summary

This study targets researchers focusing on the performance of genomic prediction models in crop breeding programs. We applied an ensemble of diverse individual genomic prediction models to predict key flowering time traits (days to anthesis and anthesis to silking interval) in two maize datasets. The ensemble approach consistently outperformed the individual models, possibly attributed to the offset of prediction errors by combining multiple different dimensions of trait genetic architecture. Improved prediction performance can lead to higher selection accuracy of desirable individuals, consequently accelerating genetic gain in crop breeding.

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; Khaipho-Burch et al. 2023; 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 for breeding applications (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 (Messina et al. 2025; 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) maize examples (Buckler et al. 2009; Chen et al. 2019). The parents of these two NAM datasets represent a

82 wide range of maize genetic diversity (Fig. 1), extending from domestication to modern inbred lines used for
83 hybrid breeding (Hufford et al. 2012). This study was oriented by two objectives. Firstly, the prediction
84 performance of the ensemble was compared with individual genomic prediction models to evaluate the
85 expectations of the Diversity Prediction Theorem for prediction performance improvement. Secondly, the
86 estimates of genomic marker effects from each genomic prediction model were compared to observe the
87 association between the diversity in quantified marker effects and the prediction performance of the ensemble.

88

89 Materials and Methods

90 1. Datasets

91 The TeoNAM dataset (Chen et al. 2019) is a collection of samples from the five recombinant inbred line (RIL)
92 populations between the maize line W22 and five teosinte types (TIL01, TIL03, TIL11 and TIL14 from *Z.*
93 *mays ssp. parviglumis* and TIL25 from *Z.mays ssp. mexicana*). In each RIL population, the F1 was backcrossed
94 with W22. After the backcross was completed, each population was self-pollinated four times to develop the
95 RIL populations. A randomised complete block design was applied to test the RILs for each population
96 contributing to the TeoNAM experiment. Each population was tested twice at the University of Wisconsin
97 West Madison Agricultural Research Station. For W22TIL01, W22TIL03 and W22TIL11, the RILs were
98 tested in the summers of 2015 and 2016. For W22TIL14, the RILs were tested in the summers of 2016 and
99 2017, and for W22TIL25, the RILs were tested in two different blocks in the summer of 2017.

100 The MaizeNAM dataset (Buckler et al. 2009) consists of samples from 25 RIL populations derived
101 from crosses between the maize line B73 and a diverse set of 25 inbred lines from temperate and tropical
102 regions. The 25 RIL populations were developed through self-pollination to the F5 generation. Each
103 population was tested in the following locations during the summer in the United States: Aurora in New York,
104 Clayton in North Carolina, Columbia in Missouri and Urbana in Illinois. Two experiments were conducted at
105 each location with a randomised design and hence each population was evaluated in the eight environments.
106 The recorded phenotype scores in each environment were leveraged to estimate the best linear unbiased
107 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 the genetic diversity that covers a continuum of maize lines starting from wild relatives to domesticated ones (Fig. 1). The TeoNAM dataset contains a high level of genetic diversity: Crosses of a temperate maize inbred with representatives of its ancestral species, Teosinte, can 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. Nucleotide diversity (π) across non-overlapping 10-kb windows was higher for the Teosinte lines used in the TeoNAM populations ($\pi = 0.0059$; *Z. mays ssp. parviglumis*), compared to the improved maize lines included as parental lines in several MaizeNAM populations ($\pi = 0.0048$) (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

We leveraged data preprocessing methods from Tomura et al. (2025b). Briefly, 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) 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 kb and a step size of 5 SNPs in both datasets.

The two flowering-related traits, days to anthesis (DTA) and anthesis to silking interval (ASI), were measured on both datasets. For the TeoNAM dataset, the DTA and ASI traits for each RIL were recorded

across two environments as reported by Chen et al. (2019). These environments were concatenated for the ensemble analysis following Tomura et al. (2025b).

3. Genetic variance component calculation

Both additive and non-additive (epistatic) variance components were estimated per population for both DTA and ASI 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).

156 4. Genomic prediction model implementation and genomic level analysis

157 Six individual genomic prediction models were implemented and their inferred trait genetic architecture was
158 extracted using the computational pipeline tool, EasiGP (Tomura et al. 2025a), in Python (v3.11). Following
159 Tomura et al (2025b), three classical genomic prediction models (ridge regression best linear unbiased
160 prediction (rrBLUP; Meuwissen et al. 2001), BayesB (Meuwissen et al. 2001) and reproducing kernel Hilbert
161 space regression (RKHS; Gianola and Van Kaam 2008)) and three machine learning models (random forest
162 (RF; Breiman 2001), support vector regression (SVR; Drucker et al. 1996) and multi-layer perceptron (MLP;
163 Rosenblatt 1958)) were applied in this study. For rrBLUP, the genomic marker effects are assumed to be
164 normally distributed for all genomic markers. In contrast, some of the genomic marker effects can be shrunk
165 to zero rather than forming a normal distribution in BayesB (Endelman 2011; de los Campos 2014). For RKHS,
166 genomic markers are mapped to the Hilbert space using a kernel that estimates the mean distance between
167 genotypes in a squared Euclidean form (del los Campos et al. 2009). RF consists of numerous decision trees
168 developed from a fraction of the given dataset. Each decision tree returns prediction values by traversing the
169 tree with input attribute values. The average of predicted values from all the decision trees is leveraged as the
170 final prediction value. SVR predicts values by optimising the position of a hyperplane that maximises the
171 number of samples included in the space (epsilon tube) between the hyperplane and the decision boundary.
172 MLP mimics human brain systems by continuous nonlinear aggregation of input data from neighbouring
173 neurons in a forward direction (Dave and Kamlesh, 2014). The ensemble-average model calculates the mean
174 predicted phenotypes from the six individual models with equal weights.

175 The following hyperparameter values were assigned to each individual genomic prediction model. For
176 rrBLUP, BayesB and RKHS, the number of iterations was set as 12,000 and the burn-in was 2,000 for the
177 three models in both datasets. Default values were leveraged for all other parameters. For RF and SVR, the
178 default hyperparameter values were leveraged for both prediction models in both datasets, except for the
179 number of trees (1,000) in RF. MLP was structured with one hidden layer in both datasets. For the TeoNAM
180 dataset, 50 neurons were included in the hidden layer with a dropout of 0 and activated by the Rectified Linear
181 Unit (ReLU; Nair et al. 2010) function. The epoch of 200 was leveraged to train the model with a learning
182 rate of 0.005 using the Adaptive Moment Estimation with Weight Decay (AdamW; Loshchilov and Hutter

183 2017) optimiser. For the MaizeNAM dataset, the number of neurons was reduced to 10 with a dropout of 0.1
184 and activated by the Rectified Linear Unit (ReLU) function. The epoch of 2,500 was used with the learning
185 rate of 0.005 using RMSprop.

186 Genomic marker effects were extracted after the development of genomic prediction models as a part
187 of the pipeline in EasiGP. For rrBLUP and BayesB, the allele substitution effects were extracted as the genomic
188 marker effects. For RF, genomic marker effects were extracted using feature importance based on the impurity-
189 based approach (Ishwaran 2015). The importance of a feature (genomic marker effect in this study) is
190 determined by the total impurity value derived from the impurity value of each decision node for the feature.
191 Higher importance is allocated to a feature if the feature can clearly divide the data points (RILs) into sub-
192 decision nodes that cluster data points with similar labels (observed phenotypes). For RKHS, SVR, and MLP,
193 Shapley scores (Shapley 1953; Lundberg and Lee 2017) were extracted as genomic marker effects. Shapley
194 scores in this study estimate the causal effect of a particular genomic marker in the form of a conditional effect.
195 Since Shapley scores are calculated cumulatively by considering possible genomic marker combinations,
196 Shapley scores were also applied to RF at the pairwise level to estimate genomic marker-by-marker interaction
197 effects. For the ensemble-average model, the mean values of the normalised genomic marker effects from all
198 the individual genomic prediction models were calculated with the same weights.

199 The extracted genomic marker effects were mapped to corresponding marker regions using circos plots
200 in the final phase of EasiGP. The generated circos plots also visualised the genome positions of known genes
201 and regulators of DTA from three previous studies. QTL regions were discovered by Chen et al. (2019) for the
202 TeoNAM dataset and Buckler et al. (2009) for the MaizeNAM dataset. Wisser et al. (2019) identified several
203 key genes affecting flowering time in maize in an empirical experiment of the short-term evolution of tropical
204 region maize lines. Lastly, Dong et al. (2012) created a list of key gene regions and their interactions for
205 flowering time in maize. They classified key gene regions into two categories (leaf and shoot apical meristem
206 (SAM)) based on the tissue or organ of the maize plant where they act to determine flowering time. These
207 genes were also classified into seven categories based on their a priori defined functions (light transduction,
208 clock, photoperiod, autonomous, aging, gibberellin and integrator).

209

5. The Diversity Prediction Theorem

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):

$$(\bar{M} - V)^2 = \sum_{i=1}^{N} \frac{(M_i - V)^2}{N} - \sum_{i=1}^{N} \frac{(M_i - \bar{M})^2}{N} \tag{4}$$

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 (4):

$$ratio = \frac{\sum_{i=1}^{N} \frac{(M_i - \bar{M})^2}{N}}{\sum_{i=1}^{N} \frac{(M_i - V)^2}{N}} \tag{5}$$

The ratio eliminates the scale difference between the trait phenotypes across prediction scenarios and hence the values of Equation (5) 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 (4)) 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.

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. 2). 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. There was a higher contribution of the non-additive variance component for ASI in both TeoNAM (0.610) and MaizeNAM datasets (0.626). Therefore, 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. 3). 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

257 ensemble-average model also had a smaller median prediction error for DTA (the ensemble-average model =
258 11.033 and the average of individual genomic prediction models = 14.699) and ASI (the ensemble-average
259 model = 4.339 and the average of individual genomic prediction models = 4.638). For the MaizeNAM dataset,
260 the median prediction accuracy for DTA and ASI was higher for the ensemble-average model (0.642 and 0.465,
261 respectively) compared to the average performance of the individual genomic prediction models (0.601 and
262 0.436, respectively). Similarly, there was improvement in the median prediction error for both traits (DTA =
263 2.389 and ASI = 0.442 in contrast to DTA = 2.978 and ASI = 0.464). The observed results indicate that the
264 ensemble-average model consistently reached a higher prediction performance compared to the average
265 performance of the individual genomic prediction models.

266

267 3. Diversity contributed to the performance improvement of the ensemble

268 The level of prediction improvement of the ensemble varied with traits and datasets, associated with the level
269 of diversity in predicted phenotypes (Table 2). The ratio defined in Equation (5) based on the Diversity
270 Prediction Theorem was higher in the TeoNAM dataset for both traits (DTA = 0.42 and ASI = 0.19) in contrast
271 to the MaizeNAM dataset (DTA = 0.33 and ASI = 0.11) (Table 2). A higher ratio value implies that the
272 predicted phenotypes from the individual genomic prediction models are more diverse, relative to the basis of
273 the mean predicted phenotypes. Since greater improvements in prediction performance were observed in the
274 TeoNAM dataset compared to the MaizeNAM dataset, the higher ratio values for both DTA and ASI in the
275 TeoNAM dataset were associated with higher diversity in the predicted phenotypes from the individual
276 genomic prediction models. Additionally, the higher ratio for DTA compared to ASI, for both datasets,
277 indicates the potential for greater prediction performance improvement for DTA than for ASI in both datasets.
278 These findings indicate that there is a positive association between the level of diversity in the predicted
279 phenotypes and the prediction performance of the ensemble.

280 The comparison of the predicted DTA at both the phenotype and genotype levels indicated a different
281 degree of information diversity between the datasets (Fig. 4). Most of the pairwise comparisons in the
282 MaizeNAM dataset demonstrated stronger positive associations (average for predicted phenotypes, $r^2 = 0.911$,

283 and average for predicted genomic marker effects, r^2 = 0.578) (Table S1) compared to the TeoNAM dataset
284 (average for predicted phenotypes, r^2 = 0.578, and average for predicted genomic marker effects, r^2 = 0.307).
285 These results indicate that, for the TeoNAM dataset, the individual genomic prediction models returned more
286 diverse predicted phenotypes by assigning different genomic marker effects. This higher diversity among the
287 individual genomic prediction models for the TeoNAM dataset likely contributed to the improved prediction
288 performance of the ensemble model in comparison to the MaizeNAM dataset (Fig. 3, 4). The relatively higher
289 diversity in the TeoNAM dataset was also observed for ASI at both the phenotype and genome levels (Fig. S1;
290 Table S2). The MaizeNAM showed a higher positive association in both levels (average for predicted
291 phenotypes, r^2 = 0.942, and average for predicted genomic marker effects, r^2 = 0.671) compared to the
292 TeoNAM dataset (average for predicted phenotypes, r^2 = 0.818, and average for predicted genomic marker
293 effects, r^2 = 0.498). The difference in r^2 between the two datasets for ASI was smaller compared to that for
294 DTA, aligning with the smaller difference in the level of prediction performance improvement by the ensemble
295 between the two datasets for ASI. Such a small level of difference also corresponded to the subtle difference
296 in the ratio of the estimated variance components between the datasets (Fig. 2). Hence, quantified genomic
297 marker effects with high diversity between models may lead to improved prediction performance of the
298 ensemble.

299

300 4. Several key gene regulators were captured by EasiGP

301 The constructed circos plots showed several overlaps between the genomic marker regions highlighted by the
302 genomic prediction models and known key gene regulators for the traits (Fig. 5). For DTA, the features of the
303 genomic prediction models were associated with several key maize flowering genes in both datasets. For
304 example, in the TeoNAM dataset, the region between 35 cM and 50 cM in chromosome 10 was repeatedly
305 targeted across genomic prediction models (with MLP as an exception). This region contains ZmCCT10,
306 involved in regulating flowering time in maize, known to upregulate the circadian clock as part of the
307 photoperiod pathway (Dong et al. 2012; Chen et al. 2019; Wisser et al. 2019). The region containing
308 ZmCCT10 also showed strong interactions with the region in chromosome 8 (between 60 cM and 80 cM).
309 This region includes ZCN8, a well-established gene controlling flowering in maize that is known to interact

310 with ZmCCT10 (Dong et al. 2012; Chen et al. 2019; Wisser et al. 2019). In the MaizeNAM dataset, in addition
311 to the same two regions identified in the TeoNAM dataset, the genomic prediction models identified another
312 region between 50 cM and 60 cM in chromosome 9 that strongly interacted with the region containing
313 ZmCCT10 in chromosome 10 and the region containing ZCN8 in chromosome 8. This chromosome 9 region
314 contains ZmCCT9 (Wisser et al. 2019), another gene involved in the photoperiod pathway that negatively
315 regulates ZCN8 (Huang et al. 2018). For ASI, several overlaps between the highlighted gene regions by the
316 genomic prediction models and QTL regions (Buckler et al. 2009; Chen et al. 2019) were also observed in
317 both datasets (Fig. S2). Such an overlap between the gene regions identified across genomic prediction models
318 and the key gene regulators from the literature in both datasets indicates that these genomic prediction models
319 reliably identified regions containing several known causative loci.

320 The comparative analysis of the extracted genomic prediction model features revealed the different
321 components of the trait genetic architecture captured by the different genomic prediction models. The pattern
322 of the genetic architecture from each individual genomic prediction model was more variable in the TeoNAM
323 dataset (Fig. 4a, S2a) in comparison to the MaizeNAM dataset (Fig. 5b, S2b), consistent with the Diversity
324 Prediction Theorem framework (Table 2, Fig. 4, S1) and the results of the ensemble model prediction. While
325 the different genomic prediction models identified many similar regions of the genome for DTA, the
326 contributions of each marker identified in these regions could vary considerably in the representation of the
327 genetic architecture in the TeoNAM dataset (Fig. 5a, S2a). In the MaizeNAM dataset, on the other hand, there
328 was more consensus across the genomic prediction models (Fig. 5b, S2b). The genomic prediction models
329 detected similar genomic regions with a stronger consensus on the contribution of each marker. It is also noted
330 that, for the MaizeNAM dataset, the genomic prediction models allocated stronger weights to the markers at
331 the peaks of the QTL identified by Buckler et al. (2009) from their analysis of the same dataset (Fig. 4b; lower
332 triangle of pairwise comparisons). In contrast, for the TeoNAM dataset, while the markers at the peaks of the
333 QTL identified by Chen et al. (2019) were included in the genomic prediction models, strong weights were
334 less frequently assigned by the genomic prediction models (Fig. 4a; lower triangle of pairwise comparisons).
335 The analysis of the genomic prediction models at the genetic level revealed that different trait genetic
336 architectures were inferred by each model.

Discussion

Breeders have continuously searched for prediction methods that can consistently and accurately distinguish individuals with the desired traits in their breeding programs (Voss-Fels et al. 2019; Crossa et al. 2025; Escamilla et al. 2025). Genomic prediction models can accurately identify these individuals under certain prediction scenarios (Azodi et al. 2019). However, a priori, it is often not possible to obtain information that determines the optimal combination of scenarios and genomic prediction models without a significant trial-and-error process (Jeon et al. 2023; Crossa et al. 2025). In contrast, the ensemble of multiple individual genomic prediction models consistently improved prediction performance in accordance with the expectations of the Diversity Prediction Theorem (Page 2018; Messina et al. 2025). The improvement in prediction performance achieved by the ensemble was marginal in some cases, depending on the diversity in the predictive information included in the individual prediction models. However, the performance of the ensemble approach was always at least comparable to 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 as 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

364 range that varies by prediction scenarios (Bian and Chen 2021; Wood et al. 2023). The benefits of such
365 constrained levels of diversity align with the implications of the theoretical framework of the Diversity
366 Prediction Theorem (Page 2018). This framework can be applied to investigate the influence of diverse
367 predicted values (Equation (4) third term) towards the prediction error of the ensemble (Equation (4) first
368 term) in relation to the average of predicted values (Equation (4) second term) from each prediction model, as
369 was conducted for the TeoNAM and MaizeNAM datasets here.

370 Given that prediction performance improvement was observed for the simplest ensemble approach
371 (averaging predicted phenotypes with equal weights), which was applied herein to demonstrate the
372 methodology, there is potential for further performance improvement using advanced ensemble design
373 algorithms. One promising approach can be weight optimisation. The minimisation of the Many-Model error
374 in the Diversity Prediction Theorem (Page 2018) can be leveraged as one method to optimise the weights,
375 since the Many-Model error represents the prediction error of the ensemble. This and other forms of ensemble
376 optimisation are an active area for further research. The leveraging of prior knowledge into ensembles is
377 another approach to potentially improve the prediction performance of ensembles. Our investigation of the
378 TeoNAM and MaizeNAM datasets captured features of several known key genes (ZmCCT9, ZmCCT10 and
379 ZCN8) and emphasised the importance of other previously unknown genomic regions that are novel regions
380 to discover new causative loci. Such extracted models of trait genetic architecture can be leveraged as prior
381 knowledge to supervise the training of genomic prediction models. Prior knowledge can provide prediction
382 models with external information to integrate natural biophysical laws within their prediction mechanisms
383 (Technow et al. 2015; Messina et al. 2018, 2025; Von Rueden et al. 2021; Cooper et al. 2025). Such
384 applications of prior knowledge can mitigate the risk of overfitting prediction models, especially when the
385 training set is small and the prediction models cannot be well-trained only using the training set (Schapire et
386 al. 2002). Hence, the prediction performance of individual genomic prediction models is expected to increase
387 and, consequently, improve the prediction performance of ensembles of these models.

388 In addition to revealing key features of the genetic architecture for traits measured in the TeoNAM and
389 MaizeNAM datasets, the ensemble of diverse genomic prediction models is demonstrated as an effective
390 approach for more consistently achieving high prediction performance across breeding scenarios. Such

391 improvements in prediction performance can be leveraged to enable the consistent identification of superior
392 individuals with the desired trait combinations for all stages of breeding programs. Future studies will explore
393 opportunities to extend the ensemble-average model used in this study, including the assessment of prediction
394 performance improvement by optimising weightings of models and integrating prior knowledge (Messina et
395 al. 2025). Overall, the ensemble of multiple genomic prediction models has the potential to improve breeding
396 programs by improving prediction accuracy to accelerate genetic gain.

397

398 Data and code availability

399 The TeoNAM dataset (Chen et al. 2019) used in this study is located at https://datacommons.cyverse.org/bro-
400 wse/iplant/home/shared/panzea/genotypes/GBS/TeosinteNAM for the genotypes and https://gsajournals.figs-
401 hare.com/articles/dataset/Supplemental_Material_for_Chen_et_al_2019/9250682 for the phenotypes. The
402 MaizeNAM dataset (Buckler et al. 2009) used in this study is located at https://cbsusrv04.tc.cornell.edu/use-
403 rs/panzea/download.aspx?filegroupid=10 for the genotypes and phenotypes. The computational tool used in
404 this study is located at https://github.com/ShunichiroT/EasiGP.

405

406 Author contribution

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

411

412 Acknowledgments

413 We thank the National Computing Infrastructure (NCI) and the Research Computing Centre (RCC) at the
414 University of Queensland for providing access to the High Performance Computing (HPC) machines. We also
415 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.

de los Campos G, Gianola D, Rosa GJ. 2009. Reproducing kernel hilbert spaces regression: a general framework for genetic evaluation. Journal of animal science. 87:1883–1887. https://doi.org/10.2527/jas.2008-1259.

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.

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.

Khaipho-Burch M, Cooper M, Crossa J, de Leon N, Holland J, Lewis R, McCouch S, Murray SC, Rabbi I, Ronald P, Ross-Ibarra J, Weigel D, Buckler ES (2023) Scale up trials to validate modified crops' benefits. Nature 621:470-473. https://doi.org/10.1038/d41586-023-02895-w.

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, Wilkinson MJ, Powell O, Cooper M. 2025a. Ensemble AnalySis with Interpretable Genomic Prediction (EasiGP): Computational Tool for Interpreting Ensembles of Genomic Prediction Models. BioRxiv (accepted in the Plant Genome). 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.

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

570 Yadav S, Wei X, Joyce P, Atkin F, Deomano E, Sun Y, Nguyen LT, Ross EM, Cavallaro T, Aitken KS *et al*.
571 2021. Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive
572 genetic effects. Theoretical and Applied Genetics. 134:2235–2252. https://doi.org/10.1007/s00122-021-
573 03822-1.

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

577 Tables and Figures

578

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

Dataset	SNPs		RIL Records
	Original	LD Filtering	
TeoNAM	11,396 - 16,110	268 - 342	438 – 616
MaizeNAM	1,106	170 - 251	126 – 196

581

582

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

587

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

588

589

590

591

592

593

594

595

596

597

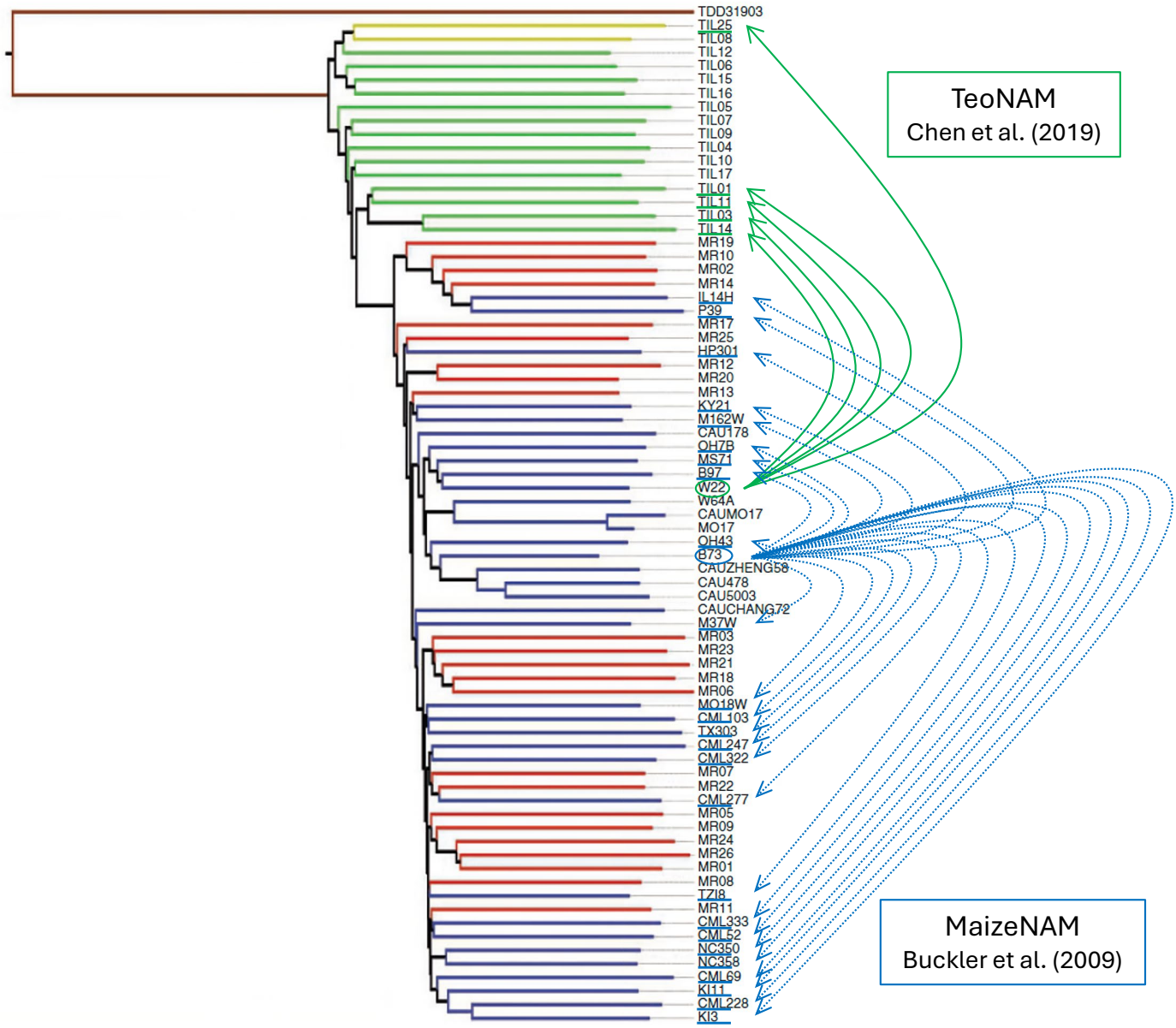

Fig. 1: Neighbour-joining tree showing sequence relationships between TeoNAM and MaizeNAM parents (adapted from Hufford et al. (2012) with permission from Professor Jeffrey Ross-Ibarra). Within the tree, each colour represents *parviglumis* (green), landraces (red), improved lines (blue), *mexicana* (yellow) and *Tripsacum* (brown), respectively. Circles around W22 and B73 represent the common parent in the TeoNAM (Chen et al. 2019) and MaizeNAM (Buckler et al. 2009) datasets, respectively. Green and blue lines under several maize line names show donor parents for the TeoNAM and MaizeNAM datasets, respectively. For the TeoNAM dataset, W22 was crossed (green arrows) with five donor parents. For the MaizeNAM dataset, B73 was crossed (blue dash arrows) with 25 donor parents.

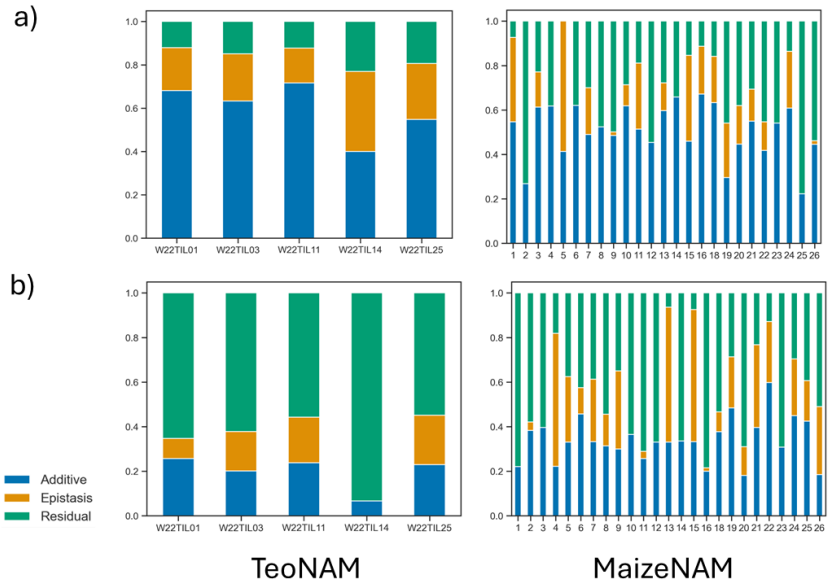

Fig. 2: The variance components of each population in the TeoNAM and MaizeNAM datasets for the a) days to anthesis (DTA) and b) anthesis to silking interval (ASI) trait. The relative magnitudes of variance components partitioned into additive (blue), epistasis (orange) and residual (green) contributions.

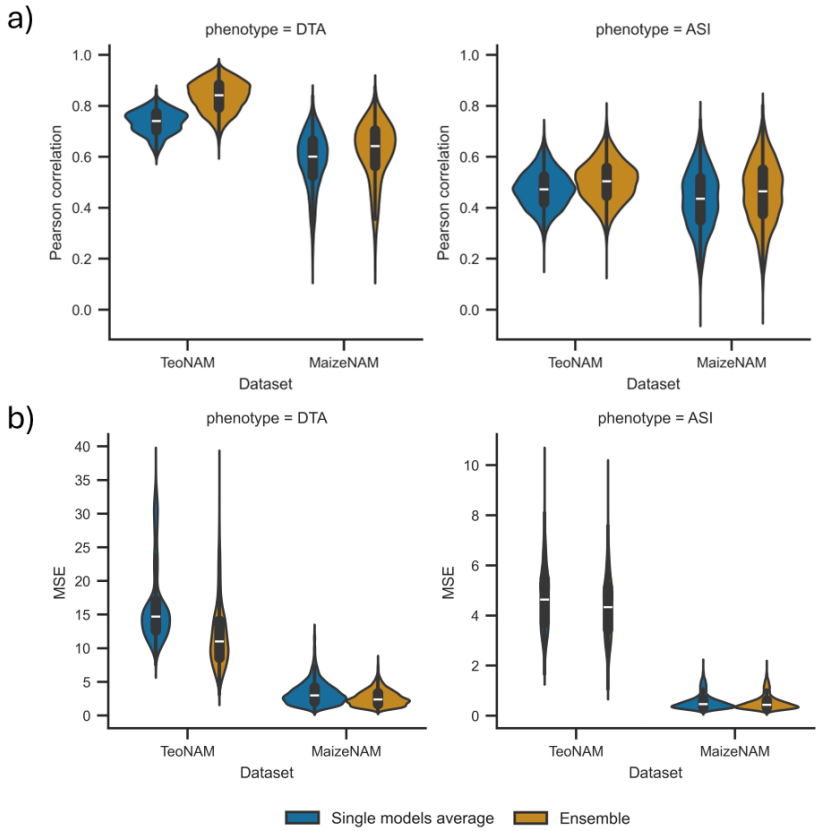

Fig. 3: The comparison of prediction performance between the average performance of the individual genomic prediction models (Single models average) and the ensemble-average model (Ensemble) for the days to anthesis (DTA) and anthesis to silking interval (ASI) traits in the TeoNAM and MaizeNAM datasets. 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.

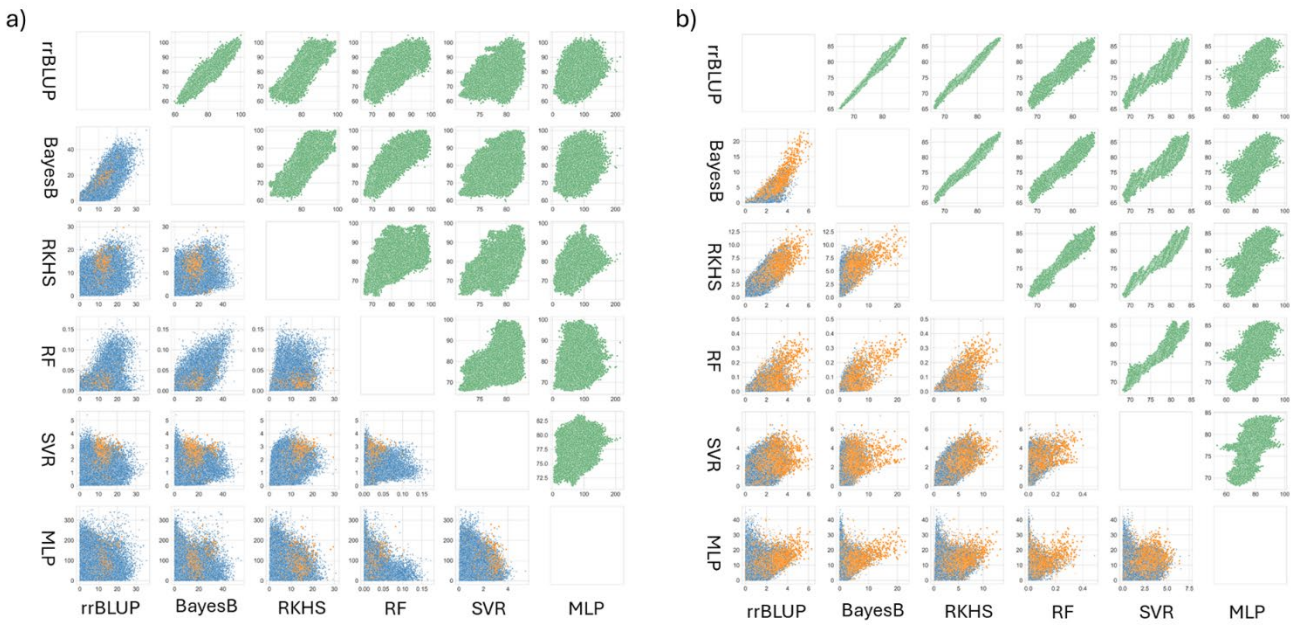

Fig. 4: Pairwise comparisons of the individual genomic prediction models for the days to anthesis (DTA) trait across all the prediction scenarios for the a) TeoNAM (7,500 prediction scenarios) and b) MaizeNAM (3,750 prediction scenarios) datasets. The prediction scenarios were generated by the combination 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.

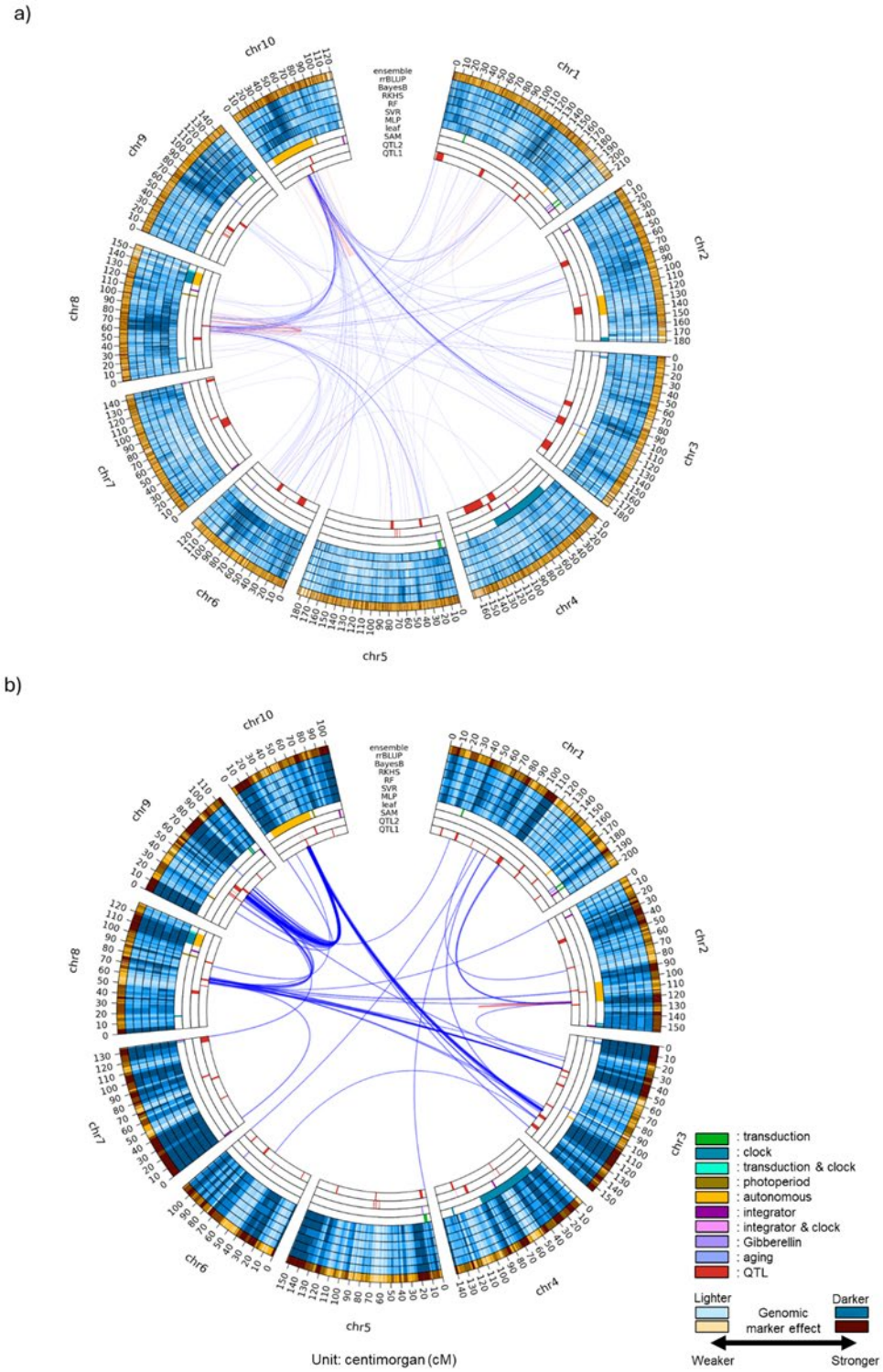

Fig. 5: Circos plots for the days to anthesis (DTA) trait using the a) TeoNAM and b) MaizeNAM datasets. The innermost (QTL 1) ring shows the QTL gene regions estimated by Chen et al. (2019) for the TeoNAM dataset and by Buckler et al. (2009) for the MaizeNAM dataset. The second innermost ring (QTL 2) 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 numbers at the outermost ring represent genetic distance in centimorgans (cM). 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 Material

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	

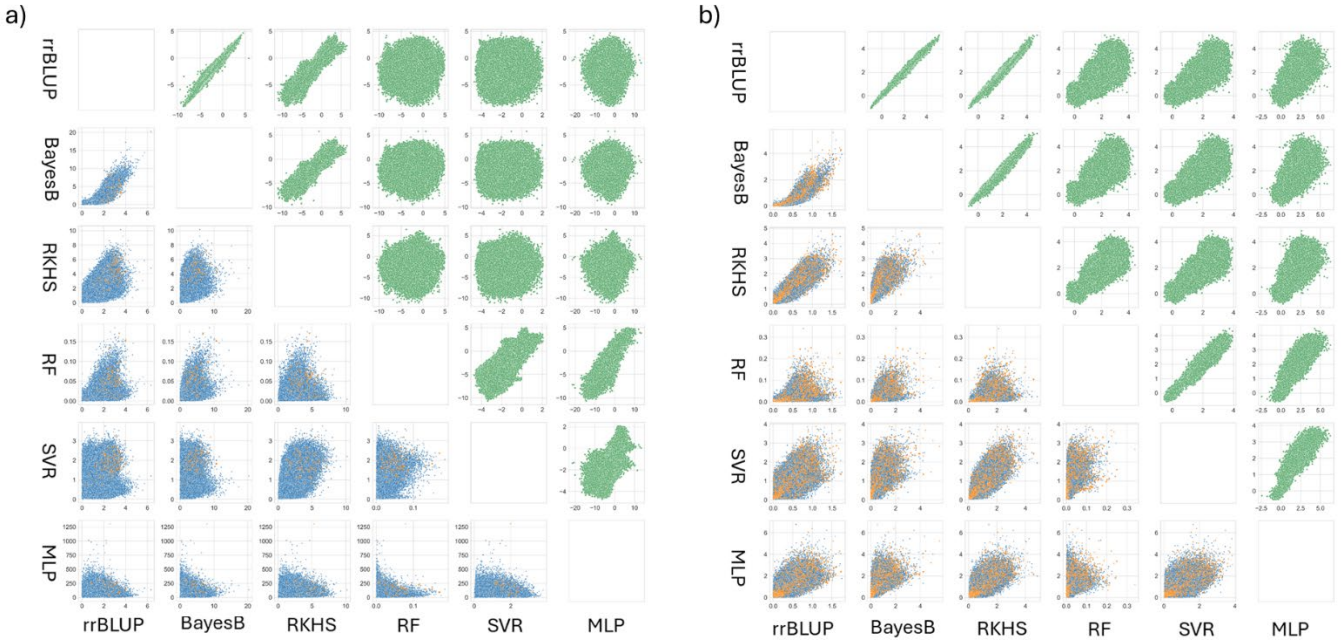

Fig. S1: Pairwise comparisons of the individual genomic prediction models for the anthesis to silking interval (ASI) trait across all the prediction scenarios for the a) TeoNAM (7,500 prediction scenarios) and b) MaizeNAM (3,750 prediction scenarios) datasets. The prediction scenarios were generated 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.

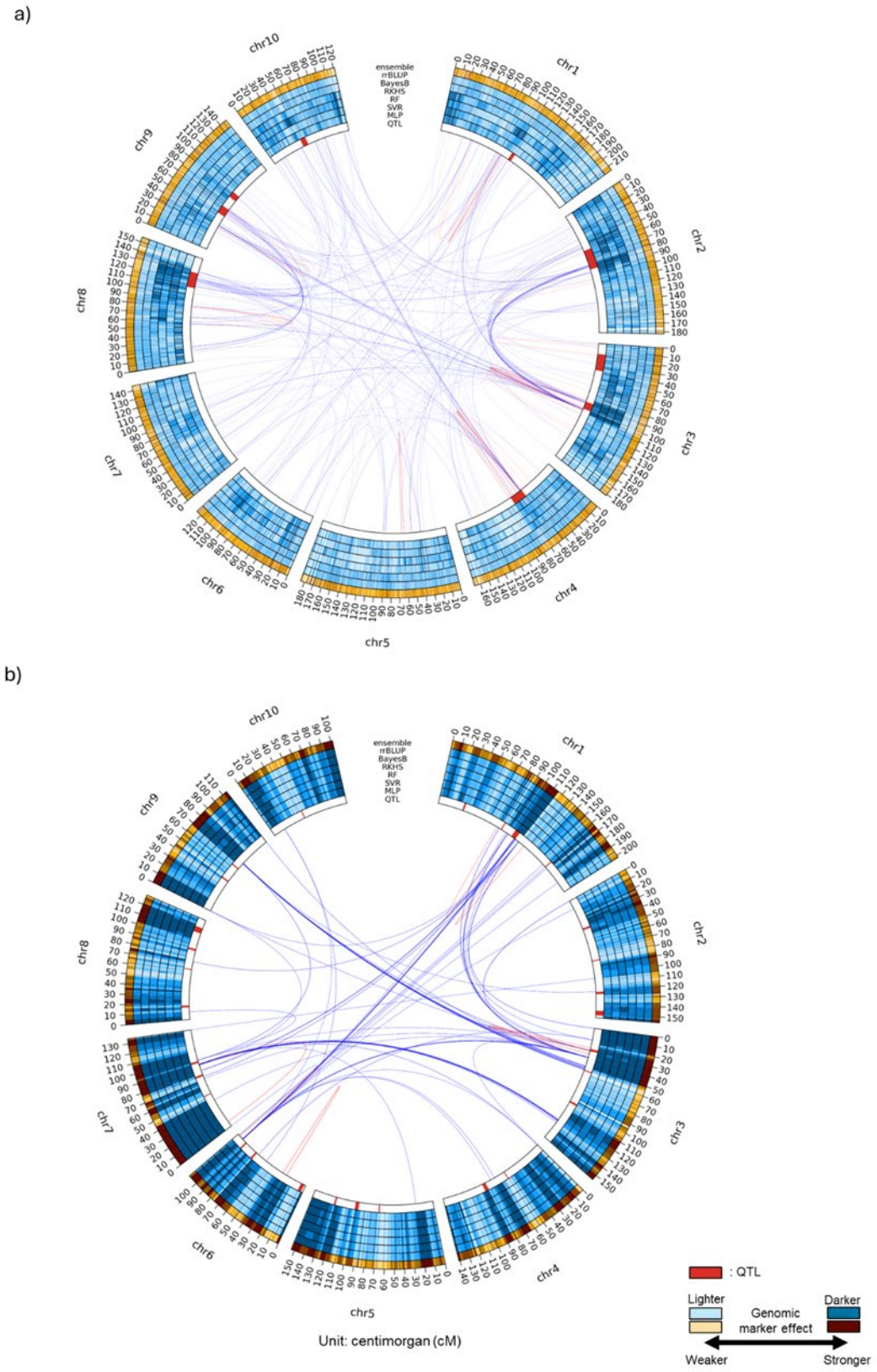

Fig. S2: Circos plots for the anthesis to silking interval (ASI) trait for the a) TeoNAM and b) MaizeNAM datasets. The innermost (QTL) ring shows the QTL gene regions estimated by Chen et al. (2019) for the TeoNAM dataset and by Buckler et al. (2009) for the MaizeNAM dataset. The blue rings (second to 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 numbers at the outermost ring represent genetic distance in centimorgans (cM). 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).