OXFORD GENETICS

GENETICS, 2024, 228(4), iyae162
https://doi.org/10.1093/genetics/iyae162
Advance Access Publication Date: 23 October 2024
Plant Genetics and Genomics

Genomic prediction of cereal crop architectural traits using models informed by gene regulatory circuitries from maize

Edoardo Bertolini ,[1] Mohith Manjunath ,[2] Weihao Ge,[2] Matthew D. Murphy,[3] Mirai Inaoka,[3] Christina Fliege,[2] Andrea L. Eveland ,[1] Alexander E. Lipka [3,*]

[1]Donald Danforth Plant Science Center, St. Louis, MO 63132, USA
[2]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[3]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*Corresponding author: Department of Crop Sciences, University of Illinois at Urbana-Champaign, W-201A Turner Hall, M/C 046, Urbana, IL 61801, USA.
Email: alipka@illinois.edu

Plant architecture is a major determinant of planting density, which enhances productivity potential for crops per unit area. Genomic prediction is well positioned to expedite genetic gain of plant architectural traits since they are typically highly heritable. Additionally, the adaptation of genomic prediction models to query predictive abilities of markers tagging certain genomic regions could shed light on the genetic architecture of these traits. Here, we leveraged transcriptional networks from a prior study that contextually described developmental progression during tassel and leaf organogenesis in maize (*Zea mays*) to inform genomic prediction models for architectural traits. Since these developmental processes underlie tassel branching and leaf angle, 2 important agronomic architectural traits, we tested whether genes prioritized from these networks quantitatively contribute to the genetic architecture of these traits. We used genomic prediction models to evaluate the ability of markers in the vicinity of prioritized network genes to predict breeding values of tassel branching and leaf angle traits for 2 diversity panels in maize and diversity panels from sorghum (*Sorghum bicolor*) and rice (*Oryza sativa*). Predictive abilities of markers near these prioritized network genes were similar to those using whole-genome marker sets. Notably, markers near highly connected transcription factors from core network motifs in maize yielded predictive abilities that were significantly greater than expected by chance in not only maize but also closely related sorghum. We expect that these highly connected regulators are key drivers of architectural variation that are conserved across closely related cereal crop species.

Keywords: genomic prediction; transcriptional networks; network motifs; maize; leaf angle; plant architecture; Plant Genetics and Genomics

Introduction

The increasingly pressing challenge of feeding the growing global population, estimated to reach 9 billion by 2050, necessitates a significant increase in food production (Hunter et al. 2017). Against the backdrop of climate change, this challenge underscores a critical need to explore cutting-edge methods for crop improvement (Mohd Saad *et al.* 2022). Plant architecture has been an important target of selection in crop improvement and central to the huge gains in productivity from breeding seen throughout the past century. With inevitable decreases in arable farmland worldwide and shifting weather patterns, modern-day crop improvement must adjust plant ideotypes for diverse environments and thus architectural traits remain a key target (Huang *et al.* 2022).

A core breeding objective to accommodate these challenges has been to increase planting density without compromising plant fitness. Consequently, modern high-density planting schemes for cereal crops have been successfully implemented, resulting in significantly increased productivity per unit area (Cao *et al.* 2022). In maize (*Zea mays*), this was realized through breeding efforts focused on architectural traits such as upright leaves, minimal branching, and tillering. Narrower leaf angle (LA) and reduced tassel branch number (TBN) not only have significant implications for crop management but also enable better light penetration into the lower canopy. This reduces competition for sunlight and hence optimizes photosynthetic efficiency. Such breeding efforts resulted in a higher number of plants per hectare and, consequently, a substantial increase in crop yields (Duvick 2005).

Plant architectural traits, including LA and inflorescence structure (e.g. TBN), exhibit high heritability across economically important crops (Upadyayula *et al.* 2006; Casa *et al.* 2008; Sary *et al.* 2022). This allows the opportunity to employ genomic prediction (GP) models that utilize signatures of genetic architecture captured by genome-wide marker sets to obtain accurate genomic estimated breeding values (GEBVs) for plants that have not been phenotyped (Bernardo 1994; Meuwissen *et al.* 2001) and therefore accelerate genetic gain. The most widely used statistical models for GP (reviewed in de Los Campos *et al.* 2013) have been shown to outperform competing marker-assisted selection approaches that only consider the effects of single, large-effect genes (Meuwissen *et al.* 2001; Bernardo and Yu 2007; Heffner *et al.* 2010). In addition to its well-established contribution to

Received on 23 May 2024; accepted on 28 September 2024

modern plant and animal breeding programs (reviewed in Van Eenennaam et al. 2014 ; Crossa et al. 2017), GP models that account for gene families or other genomic features identified in prior studies can be used to make inferences on their contribution to the overall genetic architecture of the studied trait. For example, Turner-Hissong et al. (2020) used a series of fine-tuned GP models to infer the contributions of several pathways to free amino acids composition in dry *Arabidopsis* seeds. The use of GP for making inferences on genetic architecture is crucial because approaches such as genome-wide association studies (GWAS) and quantitative trait locus (QTL) analyses typically only identify moderate–large-effect loci (Lipka et al. 2015). This misses the vast majority of the genomic signal explained by the collective effects of putatively small-effect loci underlying complex traits. Thus, a practical implication of using GP to make inferences on the genetic architecture of plant architectural traits is that it could shed important light on the contributions of genomic effects that tend to be overlooked by GWAS.

Prior studies investigated the use of certain genomic features to reduce genome-wide marker sets, e.g. accessible chromatin regions, capturing a significant portion of the phenotypic variation (Rodgers-Melnick et al. 2016; Parvathaneni et al. 2020). A recent study by Bertolini et al. (2023) demonstrated the power of fine-tuning genotype-to-phenotype models with biological information derived from transcriptional networks, which facilitated the identification of small-effect genetic loci associated with LA and TBN in maize. This approach defined transcriptional circuitries in specific developmental contexts (i.e. tassel and leaf development) (Fig. 1, left panel) by inferring (1) gene coexpression (GC) networks to identify groups of genes with similar expression patterns and (2) gene regulatory networks to predict transcription factor regulation of gene targets. This study revealed that gene sets derived from specific developmental networks could explain a significant portion of the narrow-sense heritability (h^2) for LA and TBN, which suggests these gene sets form part of the genetic architecture for these traits (Bertolini et al. 2023). Therefore, we

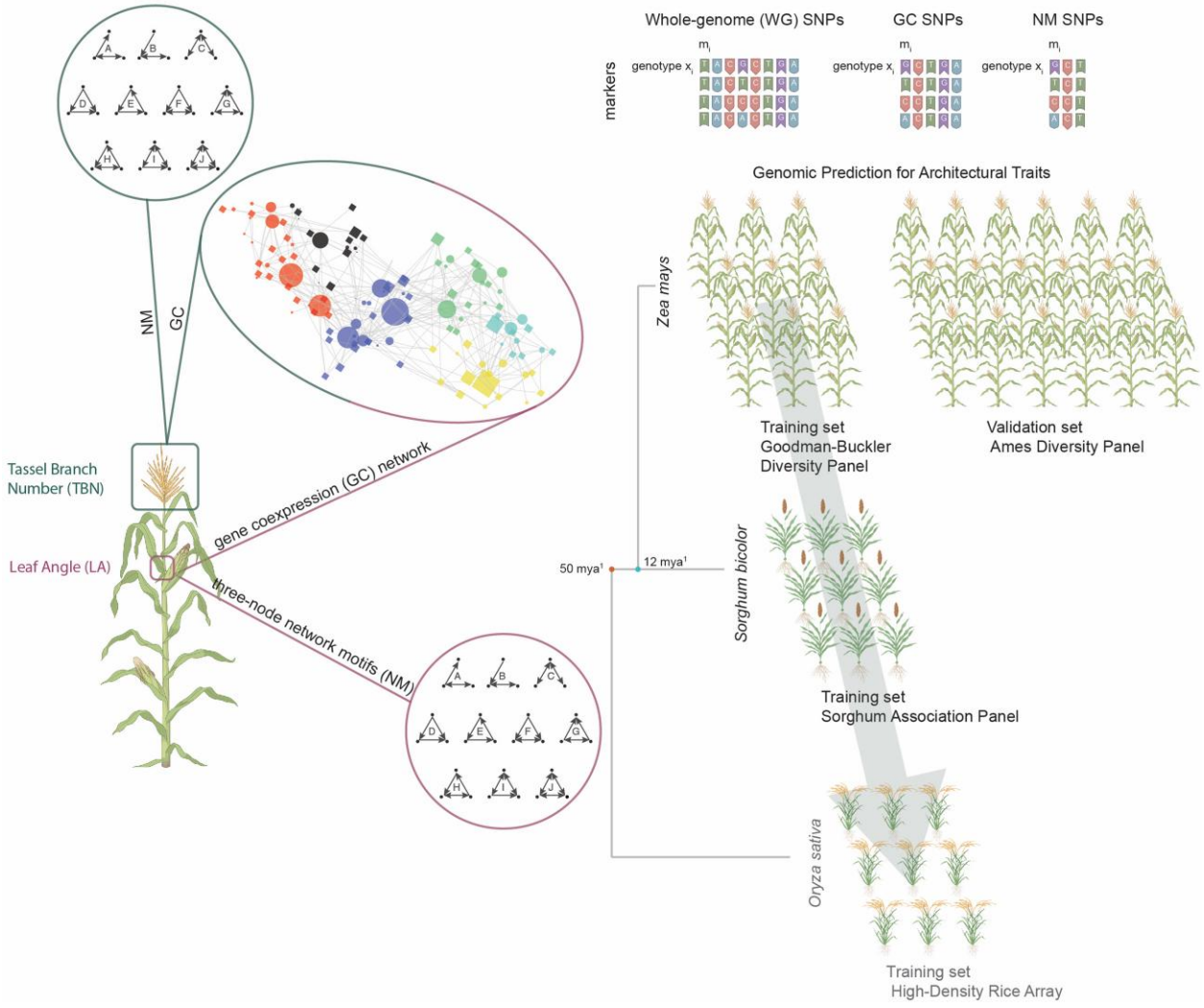

Fig. 1. Experimental scheme of GP using transcriptional networks. The left panel represents the identification of transcriptional networks from tassel primordia and the ligular region (Bertolini et al. 2023). The ball-and-stick diagram, enclosed within the ellipse, represents the 6 selected modules from the coexpression networks generated using expression data from both tissue types (Bertolini et al. 2023). The two circles represent the tissue-specific 3-node NMs with an edge number of 3 or more (from A to J). The right panel illustrates the GP approach used in this study. It includes 3 marker sets (WG, GC, and NM), each comprising markers with MAF < 0.05 and then markers with MAF > 0.05. GP models were trained in the maize Goodman–Buckler diversity panel and validated in the Ames panel. The arrow indicates the cross-species translation using markers located near sorghum and rice orthologs. Dots represent phylogenetic distance (million years ago) among the species (Swigonová et al. 2004).

expect that these sets of network genes should similarly produce reasonably high predictive abilities in GP models.

The purpose of our study was to use GP to determine the contribution of biological gene networks to the genetic architectures of LA and TBN, 2 agronomically important architectural traits in maize. We also tested the ability to translate gene network-based information from maize to other cereals including sorghum (*Sorghum bicolor*), which is closely related to maize, and rice (*Oryza sativa*), a more distantly related cereal. We considered 4 prediction scenarios (Fig. 1, right panel): (1) where we predicted GEBVs within the maize diversity panel from Flint-Garcia *et al.* (2005); (2) where the maize diversity panel in (1) was the training set and a larger maize diversity panel from Romay *et al.* (2013) was the validation set; (3) where the predictive abilities for LA using sorghum orthologs of core network genes from maize were assessed in a sorghum diversity panel (Casa *et al.* 2008); and (4) where a procedure similar to (3) was conducted on a rice diversity panel (McCouch *et al.* 2016). For all scenarios, we compared the predictive abilities of markers proximal to network gene sets to those from a genome-wide marker set, as well as an empirical distribution of prediction accuracies from randomly selected subsets of markers.

Materials and methods

Genotypic data

We analyzed 2 well-studied maize diversity panels: the Goodman–Buckler diversity panel (Flint-Garcia *et al.* 2005) and the Ames panel (also known as the North Central Regional Plant Introduction Station Panel; Romay *et al.* 2013). We also analyzed the sorghum association panel (SAP; Casa *et al.* 2008) and the high-density rice array (HDRA; McCouch *et al.* 2016). Genotyping by sequencing (GBS) data for markers segregating in both the Goodman–Buckler and Ames maize diversity panels were downloaded from Panzea (www.panzea.org) and processed following the methodology outlined in Bertolini *et al.* (2023). Specifically, genomic coordinates were uplifted to the maize reference AGPv4 (Jiao *et al.* 2017), indels and nonbiallelic markers were filtered out, and missing data were imputed using the nearest neighbor method (Money *et al.* 2015). Single-nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) < 0.01 were also discarded. The SAP GBS data (Bouchet *et al.* 2017) were downloaded from the Dryad Digital Repository (doi:10.5061/dryad.gm073). The HDRA genotypes data were downloaded from Rice Diversity (www.ricediversity.org). All genotype data were then employed to subset markers based on specific gene network subsets and different MAF cutoffs for further analysis.

Gene module information

We used transcriptional networks related to tassel branching and ligule development in maize from the Bertolini *et al.* (2023) study, including GC networks representing groups of genes with similar expression patterns and 3-node network motifs (NMs), which are elementary gene regulatory circuits of regulatory transcription factor networks. Markers within genomic coordinates of these 2 gene sets were selected based on genomic windows defined as within 2 kb from the transcription start site (TSS) and the transcription termination site (TTS) (see Bertolini *et al.* 2023 for further details). The maize GC and NM genes were translated to sorghum and rice using syntenic orthologous gene information retrieved from Zhang *et al.* (2017). Due to larger LD blocks relative to maize (Morris *et al.* 2013), the sorghum and rice genomic regions were extended by 10 kb from the TSS and TTS.

Phenotypic data

We used phenotype data for LA and TBN published in Bertolini *et al.* (2023) for our analysis. These data were from 231 lines of the Goodman–Buckler diversity panel and 1,064 lines of the Ames panel. As described in Bertolini *et al.* (2023), these data were grown in a randomized complete block design (RCBD) between 2018 and 2021. Sorghum LA phenotype data were previously collected for 296 individuals from the SAP (Casa *et al.* 2008), which were planted in a RCBD with 2 replications per location in 2010 and 2012. LA was measured from the leaf below the flag leaf, and 2 plants per replication were measured using a protractor (Mantilla Perez *et al.* 2014). Similarly, LA phenotype data were collected from a rice diversity panel of 344 varieties (Huber *et al.* 2024) using a RCBD with 4 replicates. LA was collected at an early vegetative stage, between the second and third youngest leaves and the culm.

GP model used

We employed the ridge regression best linear unbiased prediction (RR-BLUP; Whittaker *et al.* 2000; Meuwissen *et al.* 2001) model to obtain GEBVs of TBN and LA. This model equates a given trait to a linear combination of random marker effects and a random error term, as described previously (e.g. Rice and Lipka 2019), and the resulting genotype BLUPs are "shrunk" to the mean as a result of a ridge penalty (Hoerl and Kennard 1970) determined from the ratio of error variance to genetic variance. The RR-BLUP model was fitted using the rrBLUP R package (Endelman 2011).

MSTEP and USTEP models for maize

We implemented 2 multilocus stepwise model selection procedures to identify markers exhibiting strong statistical associations with TBN and LA in maize. The first procedure was the multitrait, multilocus (MSTEP) procedure, which is described in detail in Fernandes *et al.* (2022). Briefly, this procedure fits a series of multitrait, multilocus models in a stepwise manner to identify markers exhibiting strong additive associations with multiple traits. The specific markers to be included in the model are determined through a stepwise model selection procedure. In this implementation, we considered TBN and LA as the 2 response variables. The second procedure we considered was a single-trait analog of MSTEP. As done in Fernandes *et al.* (2022), we abbreviated this procedure as the univariate stepwise model selection procedure (USTEP), and we fitted it separately to TBN and then again to LA. For both of these model selection procedures, stepwise model selection was conducted in the TASSEL software (Bradbury *et al.* 2007) until a total of 10 markers were in the final models.

GPs within the Goodman–Buckler maize diversity panel

Five-fold cross-validation was performed to obtain predictive abilities for LA and TBN in the full marker set (44,930 SNPs), subsets of markers obtained from GC modules (21,362 SNPs) and from NMs (466 SNPs) from Bertolini *et al.* (2023). For each of these subsets, we also evaluated the predictive abilities of markers with MAF < 0.05 and then markers with MAF > 0.05. This was undertaken to evaluate the possibility that markers with MAF < 0.05 might capture different causal loci than markers with MAF > 0.05. Predictive abilities of models including only markers selected from MSTEP and USTEP as explanatory variables (and fitted to the appropriate training sets) were also evaluated through 5-fold cross-validation. We utilized the R packages rrBLUP (Endelman 2011) and GAPIT (Lipka *et al.* 2012) along with in-house R scripts

for GP using the RR-BLUP model. For all subsets of markers, the predictive ability was calculated as the sample mean Pearson product moment correlation coefficient (r) between observed trait values and GEBVs across all validation sets.

Using the Goodman–Buckler diversity panel to train models for GP in the Ames maize panel

We undertook another cross-validation study in which all 231 lines in the Goodman–Buckler diversity panel with available phenotypic data were used as the training set, and a subset of 1,064 lines of the Ames panel that maximize diversity in TBN and LA (Bertolini *et al.* 2023) were used as the validation set. For each of the aforementioned categories of marker sets, predictive ability was again determined by calculating the Pearson product moment correlation coefficient between observed trait values of the 1,064 lines in the Ames panel and their corresponding GEBVs.

GP within the SAP in sorghum and the HDRA in rice

A 5-fold cross-validation procedure, very similar to that described for the Goodman–Buckler maize diversity panel, was used to evaluate the predictive ability of markers in the vicinity of sorghum and rice syntenic orthologs of the GC and NM genes. For the SAP, this resulted in a total of 59,995 markers in the vicinity of the GC orthologs and 2,695 markers in the vicinity of NM orthologs. For the HDRA, we similarly obtained a total of 293,509 markers in the vicinity of GC orthologs and 10,970 markers in the vicinity of NM orthologs.

Procedure for obtaining an empirical null distribution to test for contribution of gene modules to genomic signals underlying traits

We used these GP models to make inferences on the contributions of GC and NM to the genetic architecture of TBN and LA. We followed a procedure similar to Parvathaneni *et al.* (2020) to derive an empirical distribution of prediction accuracies under the null hypothesis that the genes underlying the signals captured in the 2 gene set categories (i.e. GC and NM) are not important. For each category in each of the above GP experiments, we generated 1,000 random subsets by randomly selecting genes. Each of these subsets contained the same number of genes as those included in the selected category. Within each random subset, we then selected SNP markers using the genomic coordinates as described in the previous sections. This approach was undertaken because the maize GBS data used in this work are not able to target specific genomic regions, such as promoters and coding genes that probably contain causative variations associated with the phenotypic variability. Furthermore, given that the overarching null hypothesis focuses on gene sets, we felt that it was critical to ensure that each random subset was selected based on genes, not SNPs. We then fitted an RR-BLUP model in the respective training sets using only the SNP markers included in the random subset. Consequently, we obtained an empirical distribution of predictive abilities under this null hypothesis. The predictive ability of the given gene set was then compared with this empirical null distribution, and a P-value was subsequently calculated. Lower P-values provide stronger evidence against the null hypothesis that the tested gene set is not important. We considered statistical significance at $\alpha = 0.05$.

Results

Predictive abilities of TBN and LA traits using gene network informed marker sets were almost as high as those from whole-genome marker set

To assess the ability of the NM and GC marker sets to accurately predict GEBVs of TBN and LA for 231 accessions in the Goodman–Buckler maize diversity panel (Flint-Garcia *et al.* 2005), we conducted a 5-fold cross-validation procedure. Our results suggested that SNPs in the vicinity of GC genes had similar predictive ability to the entire whole-genome marker set (Fig. 2) with a mean predictive ability of 0.60 and 0.56 for TBN and LA, respectively. This could imply that for both traits, the variance explained by the GC marker set is similar to that of the whole-genome marker set. The predictive abilities observed in markers around GC genes were significantly greater than those derived from markers near randomly selected genes for both traits (Fig. 3). We also noted that the predictive ability of the GC set did not drop severely when using only low-MAF (MAF < 0.05) SNPs (Supplementary Table 1), potentially suggesting that both high- and low-MAF GC sets are tagging similar causal variants.

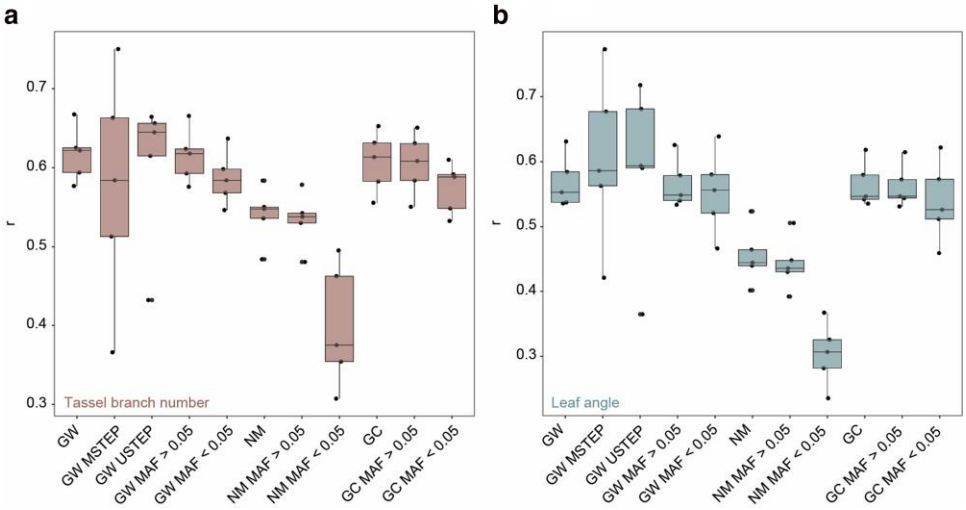

Fig. 2. Predictive ability in the Goodman–Buckler diversity panel. The image depicts the prediction accuracies in the training set. The boxplots illustrate the results of the 5-fold cross-validation tests conducted on the genome-wide (GW), the NM, and the GC marker sets at different MAF cutoffs. Y-axis represents the correlation coefficient (r) between observed trait values and GEBVs; a) and b) show results on TBN and LA, respectively.

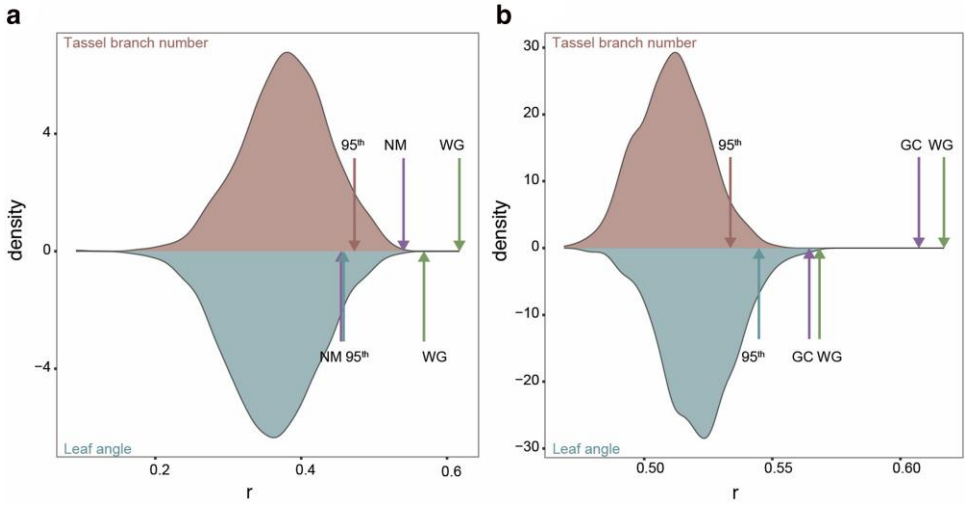

Fig. 3. Predictive performance of NM and GC marker sets in the training set. The density plots illustrate the empirical null distributions, above and below 0 density, generated based on 1,000 iterations of randomly selected genes and the subsequent selection of the colocalizing markers. Prediction results for TBN and LA are presented for the NM set a) and the GC set b). Arrows indicate the 95th percentile of the empirical null distributions, the 5-fold cross-predictive ability (average) for the NM and GC set, and the whole-genome set of markers conducted using the Goodman–Buckler diversity panel.

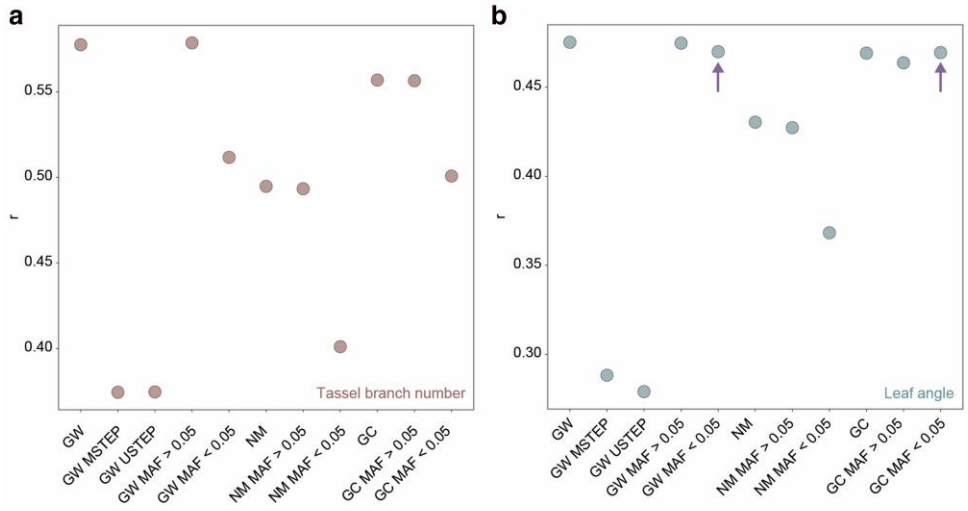

Fig. 4. Predictive ability of NM and GC marker sets in the validation set. The image represents the predictive ability results in the validation set for TBN a) and LA b) in the Ames diversity panel. Each dot represents the correlation coefficient (r) of different maker sets: genome-wide (GW), NM, and the GC markers at different MAF cutoffs.

Notably, markers associated with NM genes, despite accounting for only 10% of the whole-genome SNPs, showed only slightly lower predictive abilities for both traits (Fig. 2) relative to the whole-genome markers. When compared with the respective empirical null distribution, we observed that markers near the NM genes showed greater ability to predict TBN than was expected under the null hypothesis. In contrast, the predictability of LA by markers near the NM genes was not substantially greater than the range of predictive abilities observed across the corresponding empirical null distribution of predictive abilities (Fig. 3). For both traits, the predictive ability of the markers identified by MSTEP and USTEP were similar to those from the genome-wide marker set, highlighting a potentially strong signal derived from markers selected from these approaches. This could support previous findings, where LA was dominated by 2 major QTLs (Tian *et al.* 2011). However, the results suggested that relative to a genome-wide set of markers, MSTEP and USTEP have potential to increase the variability (and hence uncertainty) in prediction accuracies.

Markers near NM genes captured unique genomic signals in the Ames diversity panel

We next assessed the ability of GP models trained in the Goodman–Buckler diversity panel to predict GEBVs in the Ames panel. The predictive ability of GC markers was 0.55 for TBN and 0.46 for LA, while the whole-genome set achieved an accuracy of 0.57 and 0.47, respectively, for TBN and LA (Fig. 4). These relatively high predictive abilities suggest the genomic signals underlying TBN and LA in the Ames panel are similar to those underlying these 2 traits in the Goodman–Buckler panel. Interestingly, we

noted that low-MAF GC markers yielded virtually identical LA predictive abilities as the low-MAF subset of whole-genome markers, but not for TBN. We observed that the predictive abilities of markers near GC genes exceeded those of markers near NM genes in the Ames panel (Supplementary Table 1 and Fig. 4). These outcomes closely mirrored those observed during cross-validation experiments conducted within the Goodman–Buckler diversity panel (Supplementary Table 1). However, when compared with their respective empirical null distribution of predictive abilities of markers near randomly selected genes, the predictive ability of the markers near GC fell below the 95th percentile (Fig. 5a). In comparison, the predictive ability of NM genes was substantially higher than what would be expected under the null hypothesis that the NM genes are not making a meaningful contribution to the genomic signal of TBN and LA (Fig. 5b). Finally, we observed that the predictive abilities of the markers identified from MSTEP and USTEP in the Goodman–Buckler diversity panel were notably lower than those for any other considered subset of markers.

Markers near NM orthologs in sorghum showed higher predictive abilities for LA than expected by chance

To test whether context-specific biological data from maize could be used for accurately predicting parallel phenotypes in sorghum, a closely related cereal crop, we used the GC and NM genes in maize to predict LA in sorghum. We observed that the ability of markers proximal to sorghum syntenic orthologs of GC genes to

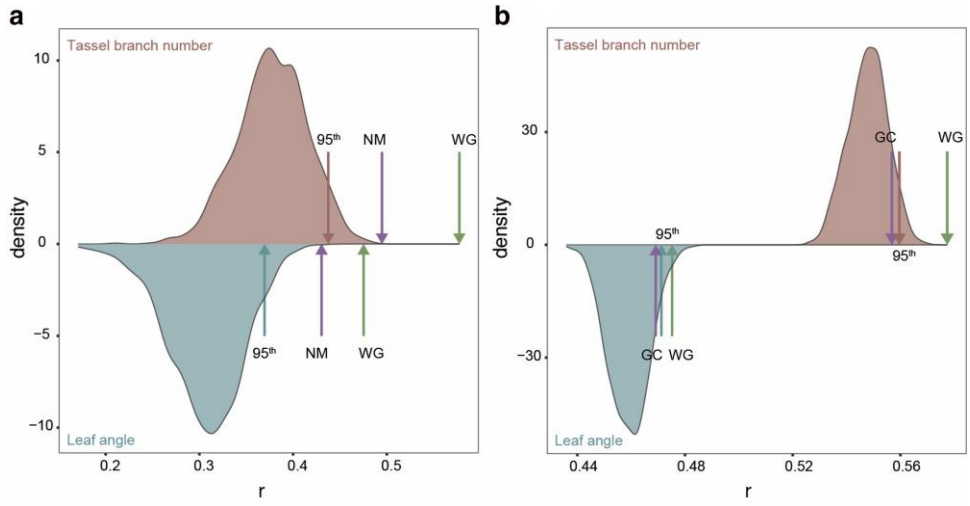

Fig. 5. Predictive performance of NM and GC markers in the validation set. The density plots illustrate the empirical null distributions, above and below 0 density, generated based on 1,000 iterations of randomly selected genes and the subsequent selection of the colocalizing markers. Prediction results for TBN and LA are presented for the NM set a) and the GC set b). Arrows indicate the 95th percentile of the empirical null distributions, the predictive ability for the NM and GC set, and the whole-genome set of markers conducted using the Ames diversity panel. Δ represents the difference in predictive ability between the WG and the marker subsets, NM or GC.

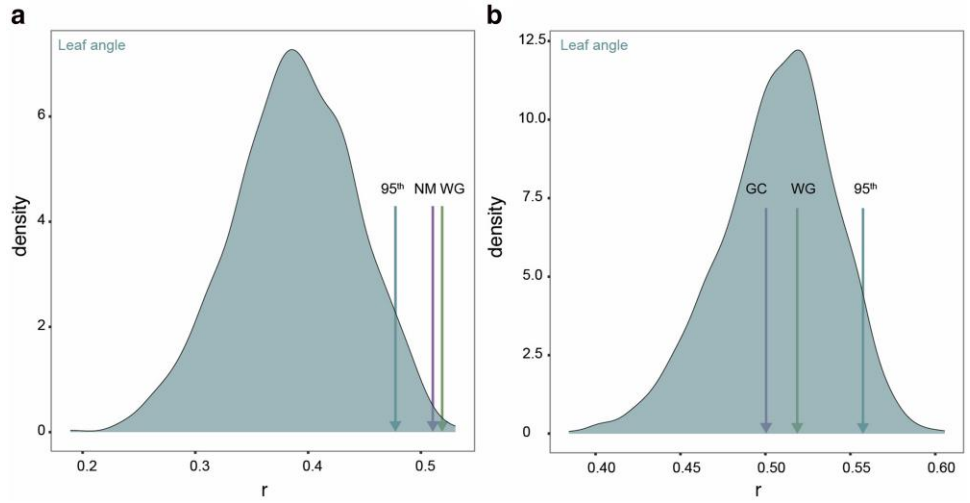

Fig. 6. Cross-species predictive ability between maize and sorghum. The density plots illustrate the empirical null distributions generated based on 1,000 iterations of randomly selected sorghum genes and the subsequent selection of the colocalizing markers. LA prediction results are represented for the NM set a) and the GC set b). Arrows indicate the 95th percentile of the empirical null distributions, the 5-fold cross-predictive ability (average) for the NM and GC set and the whole-genome set of markers conducted using the SAP.

predict GEBVs of LA was similar to those in the vicinity of orthologous NM genes (Supplementary Table 2 and Fig. 6), with a mean predictive ability of 0.50 and 0.51 for GC and NM, respectively. This similarity suggests that the GC and NM ortholog marker sets might be tagging similar subsets underlying causal variants of LA in sorghum. However, we also noted that the empirical null distribution of predictive abilities from randomly selected genes corresponding to the GC set tended to be larger than a comparable distribution corresponding to the NM set (Fig. 6). These results suggest that markers associated with NM genes play a more substantial role in the genomic signal underlying LA variance in the SAP. These overall findings closely match those from the analysis of the maize lines in the Ames diversity panel.

Cross-species predictive ability for LA between maize and rice was not statistically significant

We further assessed the cross-species translatability of markers associated with network genes and their applicability between more distantly related grass species, i.e. maize and rice. As done in sorghum, we predicted GEBVs of rice LA based on SNPs near rice syntenic orthologs of GC and NM genes. The whole-genome marker set yielded a mean predictive ability of 0.27. Given the low trait heritability ($h^2 = 0.32$; see Huber *et al.* 2024), this level of prediction was expected. However, the predictive abilities of the GC and NM sets tended to be less than those from selected markers used to generate the null distribution (Supplementary Table 3 and Fig. 7).

Discussion

We assessed the ability of markers located within and around 2 biological network-informed gene sets from maize to predict breeding values of plant architectural traits in 3 agronomically important crops. We observed that the set of transcription factor-encoding genes associated with recurrent NMs gave higher predictive abilities in maize and sorghum than expected by chance, but not in rice. This suggests that regulatory networks derived from 1 species (i.e. maize) can be used to inform loci contributing to the genetic architecture in a closely related species (sorghum). Our results also showed that this did not hold up when translating to rice, a more distantly related species; however, there are other factors that may have confounded this analysis, as described below.

NMs can capture information underlying the genetic architecture of LA important for cross-species inferences

Our results support our hypothesis on the contributions of context-specific gene regulatory networks to the genetic architecture of LA and TBN. This suggests that finely tuned GP models including only markers in the vicinity of NM genes can effectively infer elements of the genetic architecture of complex traits. Results from our cross-species analyses are likely attributed to a shared set of functionally constrained regulators that play important roles in the genotype–phenotype relationship underlying LA genetic architecture in maize and sorghum, but not in rice. This could be indicative of the close evolutionary distance between maize and sorghum (Wang *et al.* 2015) and aligns with prior findings showing gene regulatory conservation among syntenic orthologs in these species (Zhang *et al.* 2017).

However, in the case of rice, we must account for differences in the developmental stage at the time of phenotype collection (these were collected at an early vegetative growth stage), enhanced tillering compared with the other species, and the different methods used for quantifying LA in the rice data set (Huber *et al.* 2024). All of these factors may have confounded our results. Ideally, GP models in rice should be trained on LA data from mature rice plants as was done for maize and sorghum. Accounting for these differences could rule out the possibility that the observed low prediction accuracies for markers near synthetic rice orthologs of maize NM features arose because the key transcription factors underlying LA in rice change across growth stages.

GC networks sufficiently capture meaningful contributions to genetic architecture within the panel where the traits were quantified

In contrast to our results with the NM gene sets, the only situation where we saw evidence of markers surrounding GC network genes yielding higher prediction accuracies than expected by chance was within the Goodman–Buckler diversity panel. This suggests that the causal variability captured by the GC networks could be very specific to the data set being analyzed, and is potentially prone to overfitting to such an extent that they cannot capture signatures of genetic architecture, even for different panels within

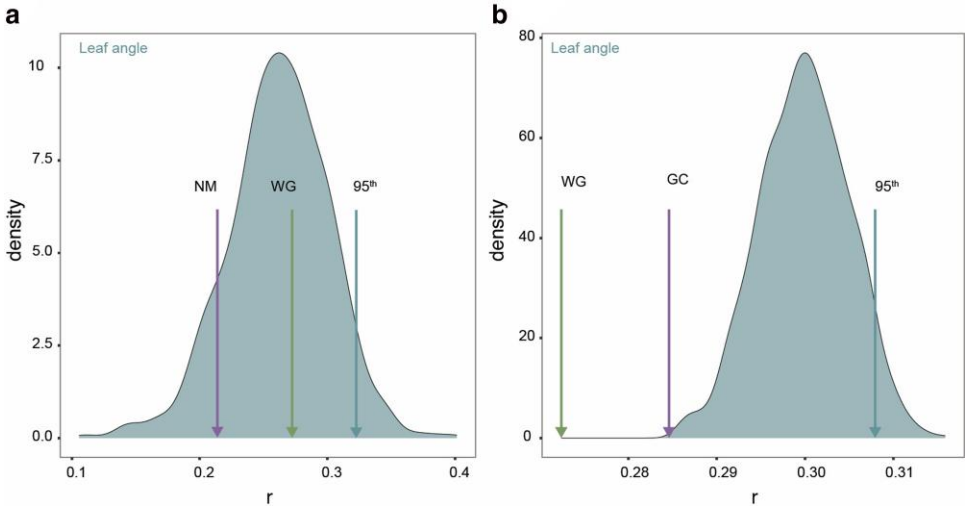

Fig. 7. Cross-species predictive ability between maize and rice. The density plots illustrate the empirical null distributions generated based on 1,000 iterations of randomly selected rice genes and the subsequent selection of the colocalizing markers. LA prediction results are represented for the NM set a) and the GC set b). Arrows indicate the 95th percentile of the empirical null distributions, the 5-fold cross-predictive ability (average) for the NM and GC set, and the whole-genome set of markers conducted using the HDRA panel.

the same species. More broadly, this result implies that using correlations between gene expressions is insufficient for capturing markers that have high accuracy when predicting unrelated individuals. This could reflect the coexpression network rewiring influenced by specific selective pressures. NMs instead capture building block patterns within the complex networks of recurrent transcription factors that might preserve functional conservation intra/interspecies. Therefore, the identification of recurring transcription factors associated with 3-node NMs helps prioritize genes that are more likely to serve as key regulators, as well as subset markers linked to genes that contribute to the genetic architecture in unrelated individuals and environments.

Using predictive abilities from GP to infer genetic architecture

This study demonstrated the potential for using GP to make inferences on genetic architecture. In practice, GP is almost overwhelmingly used to predict GEBVs of crops or livestock using whole-genome marker data, as opposed to inferring which genomic regions are likely to contain features that biologically control trait variability (Rice and Lipka 2021). Indeed, the application of GP to make such inferences should be discouraged because the number of markers in a typical data set vastly exceeds the number of individuals (see de Los Campos *et al.* 2013 for an overview of GP models). This "p >> n" scenario leads to the priors and/or penalties used in GP having such a major influence on marker effect estimates that different penalties and/or priors could identify different regions of strong statistical associations for the same trait (see e.g. Gianola 2013 for an in-depth description). Similar to other studies (e.g. Turner-Hissong *et al.* 2020), we circumvented this problem by comparing the predictive abilities of several GP models, each that focus on biologically informed subsets of the genome. Given the general similarities in predictive abilities between the whole-genome sets and the GC and NM sets, our work has far greater potential to facilitate inferences on basic biology than to change the justifiably accepted use of genome-wide markers to predict GEBVs. Nevertheless, follow-up studies should be conducted to determine the extent to which the proportions of trait variance explained by both of the marker sets differ from the whole-genome marker sets for a wider set of plant architectural traits. If these follow-up studies confirm our findings on the predictive abilities of the NM sets, they would underscore that substantial insight into genomic architecture can be made by fitting off-the-shelf GP models to a priori biologically informed marker subsets.

This work also highlighted how running existing GP models on subsets of markers can be used to compare and contrast genetic architecture between 2 traits within the same species. For instance, Figs. 2 and 4 show that there are differences in predictive ability across LA and TBN in both of the maize panels. These differences could highlight specific areas of genetic architecture that are distinct for the 2 traits. Conversely, Figs. 2 and 4 also highlight areas where the genetic architectures are comparable between LA and TBN; for example, the contributions of GC to the overall genetic architecture appear to be similar for LA and TBN within each panel. In general, the ability to make such inferences suggests that it is possible to gain insight into contrasting features of genetic architecture between traits by comparing predictive abilities of marker subsets near genomic features identified in a priori studies.

Conclusion

We used an innovative GP approach informed by gene regulatory circuitries to study the genetic architecture of complex traits. Our analyses suggest that NM facilitates the translation of biological information related to plant architecture across different diversity panels within a species, as well as between closely related species, as illustrated for maize and sorghum. This suggestive convergence of functionally constrained regulators underlying plant architectural traits opens up promising avenues for targeted breeding practices for both maize and sorghum, which can lead to optimized plant architecture for high-density planting and enhanced agricultural productivity.

Data availability

Scripts used to conduct GP and all data generated in this study are publicly available and archived online for download at https://doi.org/10.6084/m9.figshare.26733742.v1.

Supplemental material available at GENETICS online.

Acknowledgments

We acknowledge the National Center of Supercomputing Applications at UIUC for the computational facilities that made it possible to conduct the analyses in a timely manner.

Funding

This research is funded by the National Science Foundation Plant Genome Research Project award #IOS-1733606 to ALE and AEL.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Bernardo R. 1994. Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34(1):20–25. doi:10.2135/cropsci1994.0011183X003400010003x.

Bernardo R, Yu J. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47(3):1082–1090. doi:10.2135/cropsci2006.11.0690.

Bertolini E, Rice BR, Braud M, Yang J, Hake S, Strable J, Lipka AE, Eveland AL. 2023. Regulatory variation controlling architectural pleiotropy in maize. bioRxiv 553731. https://doi.org/10.1101/2023.08.19.553731 , preprint: not peer reviewed.

Bouchet S, Olatoye MO, Marla SR, Perumal R, Tesso T, Yu J, Tuinstra M, Morris GP. 2017. Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. Genetics. 206(2):573–585. doi:10.1534/genetics.116.198499.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 23(19):2633–2635. doi:10.1093/bioinformatics/btm308.

Cao Y, Zhong Z, Wang H, Shen R. 2022. Leaf angle: a target of genetic improvement in cereal crops tailored for high-density planting. Plant Biotechnol J. 20(3):426–436. doi:10.1111/pbi.13780.

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S. 2008. Community resources and strategies for association mapping in sorghum. Crop Sci. 48(1):30–40. doi:10.2135/cropsci2007.02.0080.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, *et al.* 2017. Genomic selection in plant

breeding: methods, models, and perspectives. Trends Plant Sci. 22(11):961–975. doi:10.1016/j.tplants.2017.08.011.
de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 193(2):327–345. doi:10.1534/genetics.112.143313.
Duvick DN. 2005. Genetic progress in yield of United States maize *Zea mays* L. Maydica. 50:193–202.
Endelman JB. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 4(3):250–255. doi:10.3835/plantgenome2011.08.0024.
Fernandes SB, Casstevens TM, Bradbury PJ, Lipka AE. 2022. A multi-trait multi-locus stepwise approach for conducting GWAS on correlated traits. Plant Genome. 15(2):e20200. doi:10.1002/tpg2.20200.
Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES. 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44(6):1054–1064. doi:10.1111/j.1365-313X.2005.02591.x.
Gianola D. 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 194(3):573–596. doi:10.1534/genetics.113.151753.
Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME. 2010. Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50(5):1681–1690. doi:10.2135/cropsci2009.11.0662.
Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 12(1):55–67. doi:10.1080/00401706.1970.10488634.
Huang X, Huang S, Han B, Li J. 2022. The integrated genomics of crop domestication and breeding. Cell. 185(15):2828–2839. doi:10.1016/j.cell.2022.04.036.
Huber M, Julkowska MM, Snoek LB, van Veen H, Toulotte J, Kumar V, Kajala K, Sasidharan R, Pierik R. 2024. Towards increased shading capacity: a combined phenotypic and genetic analysis of rice shoot architecture. Plants People Planet. 6(1):128–147. doi:10.1002/ppp3.10419.
Hunter MC, Smith RG, Schipanski ME, Atwood LW, Mortensen DA. 2017. Agriculture in 2050: recalibrating targets for sustainable intensification. Bioscience. 67:386–391. https://doi.org/10.1093/biosci/bix010.
Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang Bo, Campbell MS, Stein JC, Wei X, Chin C-S, *et al.* 2017. Improved maize reference genome with single-molecule technologies. Nature. 546(7659):524–527. doi:10.1038/nature22971.
Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, Gore MA. 2015. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol. 24:110–118. doi:10.1016/j.pbi.2015.02.010.
Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics. 28(18):2397–2399. doi:10.1093/bioinformatics/bts444.
Mantilla Perez MB, Zhao J, Yin Y, Hu J, Salas Fernandez MG. 2014. Association mapping of brassinosteroid candidate genes and plant architecture in a diverse panel of *Sorghum bicolor*. Theor Appl Genet. 127(12):2645–2662. doi:10.1007/s00122-014-2405-9.
McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, *et al.* 2016. Open access resources for genome-wide association mapping in rice. Nat Commun. 7(1):10532. doi:10.1038/ncomms10532.
Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157(4):1819–1829. doi:10.1093/genetics/157.4.1819.
Mohd Saad NS, Neik TX, Thomas WJW, Amas JC, Cantila AY, Craig RJ, Edwards D, Batley J. 2022. Advancing designer crops for climate resilience through an integrated genomics approach. Curr Opin Plant Biol. 67:102220. doi:10.1016/j.pbi.2022.102220.
Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. 2015. LinkImpute: fast and accurate genotype imputation for nonmodel organisms. G3 (Bethesda). 5(11):2383–2390. doi:10.1534/g3.115.021667.
Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, *et al.* 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A. 110(2):453–458. doi:10.1073/pnas.1215985110.
Parvathaneni RK, Bertolini E, Shamimuzzaman M, Vera DL, Lung P-Y, Rice BR, Zhang J, Brown PJ, Lipka AE, Bass HW, *et al.* 2020. The regulatory landscape of early maize inflorescence development. Genome Biol. 21(1):165. doi:10.1186/s13059-020-02070-8.
Rice B, Lipka AE. 2019. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. Plant Genome 12(1):180052. doi:10.3835/plantgenome2018.07.0052.
Rice BR, Lipka AE. 2021. Diversifying maize genomic selection models. Molec Breed. 41(5):33. doi:10.1007/s11032-021-01221-4.
Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. Proc Natl Acad Sci U S A. 113(22):E3177–E3184. doi:10.1073/pnas.1525244113.
Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, *et al.* 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14(6):R55. doi:10.1186/gb-2013-14-6-r55.
Sary DN, Badriyah L, Sihombing RD, Syauqy TA, Mustikarini ED, Prayoga GI, Santi R, Waluyo B. 2022. Estimation of heritability and association analysis of agronomic traits contributing to yield on upland rice (*Oryza sativa* L.). Plant Breed Biotechnol. 10(4):232–243. doi:10.9787/PBB.2022.10.4.232.
Swigonová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of sorghum and maize genome progenitors. Genome Res. 14(10a):1916–1923. doi:10.1101/gr.2332504.
Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet. 43(2):159–162. doi:10.1038/ng.746.
Turner-Hissong SD, Mabry ME, Beissinger TM, Ross-Ibarra J, Pires JC. 2020. Evolutionary insights into plant breeding. Curr Opin Plant Biol. 54:93–100. doi:10.1016/j.pbi.2020.03.003.
Upadyayula N, da Silva HS, Bohn MO, Rocheford TR. 2006. Genetic and QTL analysis of maize tassel and ear inflorescence architecture. Theor Appl Genet. 112(4):592–606. doi:10.1007/s00122-005-0133-x.
Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM. 2014. Applied animal genomics: results from the field. Annu Rev Anim Biosci. 2(1):105–139. doi:10.1146/annurev-animal-022513-114119.
Wang X, Wang J, Jin D, Guo H, Lee T-H, Liu T, Paterson AH. 2015. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. Mol Plant. 8(6):885–898. doi:10.1016/j.molp.2015.04.004.
Whittaker JC, Thompson R, Denham MC. 2000. Marker-assisted selection using ridge regression. Genet Res. 75(2):249–252. doi:10.1017/S0016672399004462.
Zhang Y, Ngu DW, Carvalho D, Liang Z, Qiu Y, Roston RL, Schnable JC. 2017. Differentially regulated orthologs in sorghum and the subgenomes of maize. Plant Cell. 29(8):1938–1951. doi:10.1105/tpc.17.00354.

Editor: J. Endelman