

Uncovering genetic mechanisms underlying trait variation in switchgrass using explainable artificial intelligence

Paulo Izquierdo^{1,2}, Xiaoyu Weng³, Thomas E. Juenger³, Jason Bonnette³, Yuko Yoshinaga⁴, Chris Daum⁴, Anna Lipzen⁴, Kerrie Barry⁴, Matthew Blow⁴, Melissa D. Lehti-Shiu², David B. Lowry^{1,2,5*}, Shin-Han Shiu^{1,2,6,*}

¹DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824, USA

²Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

³Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA

⁴U.S. Department of Energy Joint Genome Institute (JGI), Lawrence Berkeley National Laboratory, Berkeley, California, USA

⁵Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA

⁶Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA

*: Corresponding authors

Abstract

Uncovering the genetic architecture of quantitative traits is challenging because polygenic control yields small individual gene effects and because gene–gene and genotype-by-environment interactions add further complexity. To understand the genetic basis of polygenic traits and their plasticity across environments, we integrated genome-wide SNPs and RNA-seq transcript data with interpretable statistical and machine learning models in a switchgrass (*Panicum virgatum*) diversity panel grown at contrasting field sites in Michigan and Texas. Notably, in addition to single environments, our trait prediction models were able to predict phenotypic differences, across environments i.e., plasticity. By interpreting trait prediction models with explainable artificial intelligence methods, we identified important features—genes that are the most predictive of flowering time and annual biomass production across environments, based on their associated gene expression levels and nearby SNPs. This approach recovered canonical flowering regulators and revealed novel, environment-specific candidate flowering genes. Further, transcriptome models consistently recovered more switchgrass genes homologous to experimentally validated genes in Arabidopsis and rice than SNP-based models. Feature interaction scores from the models also allow the identification of trait- and environment-dependent gene–gene interactions, where flowering time showed stronger and more abundant interactions than biomass. While some of the interactions identified are consistent with the link between flowering time and yield, most are novel predictors that need to be further evaluated. Together, these results demonstrate that interpretable genomic prediction with explainable artificial intelligence approaches can convert trait prediction models into mechanistic hypotheses about putative causal genes and interactions controlling traits within and across environments. These results will help to prioritize target genes for validation and inform germplasm selection for cultivar improvement.

Introduction

Polygenic traits are controlled by multiple loci with varying effects. The contributions of these loci are further modulated by interactions with other genes (i.e., gene–gene [G×G]

or epistatic interactions) and the environment (i.e., genotype-by-environment [G×E interactions), complicating genetic mapping and the identification of loci underlying traits. Understanding the genetic mechanisms of G×E is important because it represents a key source of genetic variation that can be leveraged for climate-resilient breeding . For example, G×E can reveal conditionally neutral loci that confer trait advantages in specific environments without incurring trade-offs in others¹, offering a pathway to harness adaptive plasticity in crops. However, despite a large body of research on trait architecture, most studies have been conducted in single environments. As a result, both the contributions of G×E interactions to trait variation and their underlying genetic mechanisms remain largely unresolved^{2,3}.

To dissect the genetic mechanisms that govern trait variation within environments and the plasticity arising from G×E, it is advantageous to study species spanning broad environmental gradients⁴. One such species is switchgrass (*Panicum virgatum*), a native North American perennial grass with a broad latitudinal range and pronounced upland, lowland, and coastal ecotypic differentiation^{5,6}. This distribution has enabled the establishment of common-garden diversity panels across a latitudinal gradient, which have revealed climate adaptation among genotypes⁶. Despite the major phenotypic differences between ecotypes, earlier findings revealed that allelic variation within ecotypes could either increase or decrease traits such as biomass, depending on the environment⁷.

In addition to the availability of genetic resources facilitating studies of G×E, recent advances in high-throughput genotyping, transcriptomics, and data-driven modeling are opening new opportunities to connect predictions with underlying biological mechanisms

across multiple environments^{8,9}. Integrating omics data with interpretable machine learning, or explainable artificial intelligence¹⁰, has recovered benchmark genes and revealed novel candidates for flowering-time control in maize (*Zea mays*) and *Arabidopsis thaliana* (Azodi et al., 2020; Wang et al., 2024). Moreover, it is possible to generate genetic interaction predictions as part of model interpretation using explainable Artificial Intelligence tools like SHAP (SHapley Additive exPlanations)¹¹. SHAP quantifies the contribution of each feature to individual predictions in a consistent and additive manner, enabling precise interpretation of complex, non-linear models. These predictions have provided modeling-based hypotheses for epistasis in both humans and *Arabidopsis*, some of which are supported by experimental evidence^{12,13}. Together, these approaches offer a framework for mechanistic discovery across environments.

Despite these advances, we still lack a clear understanding of how genomic and transcriptomic signals contribute differently to complex trait variation and plasticity across environments. It remains unresolved whether transcriptome-based models capture distinct, environmentally responsive information compared to genotype-based models. Moreover, we do not yet know how gene effects are modulated by genetic background and environment, or how gene–gene interactions change across conditions. Addressing these gaps is critical for translating predictive models into biological hypotheses about trait regulation and plasticity.

To address these gaps, we leveraged phenotypic data from a switchgrass diversity panel grown in common gardens in Texas (TX) and Michigan (MI) to dissect the genetic basis of flowering time and biomass in two contrasting environments. We generated

machine learning models to predict trait variation within each site and plasticity between sites using single nucleotide polymorphisms (SNPs) identified from existing whole-genome sequencing data⁶ and newly generated transcriptomic variation data. To interpret the models, we identified genes and gene-gene interactions important for trait predictions and compared them to genes homologous to experimentally validated genes associated with flowering and biomass in Arabidopsis and rice. Finally, we evaluated how gene effects are influenced by genetic background and environmental context, and how genes contribute to trait plasticity across environments.

Results and discussion

Environmental variation drives phenotypic and transcriptional plasticity in switchgrass

To investigate the extent to which traits are influenced by environmental variation, we measured six traits (green-up, emergence, flowering time, tiller count, panicle height, biomass) for 462 tetraploid genotypes from a switchgrass diversity panel grown in Texas (TX) and Michigan (MI), which differ markedly in latitude (**see Methods**). Phenotypic data were collected at both locations during the 2021 growing season. To quantify phenotypic plasticity across environments, we calculated Pearson correlation coefficients (PCCs) of trait values across environments. PCC values near ± 1 indicate low plasticity (i.e., consistent trait rankings across environments and little G×E), and values near 0 indicate high plasticity (i.e., strong G×E effects)¹⁴. PCCs ranged from –0.13 ($p < 8.3 \times 10^{-3}$) for green-up to 0.75 ($p < 2.2 \times 10^{-16}$) for flowering time, indicating variable plasticity across traits (**Supplementary Fig. S1**). Next, we evaluated the extent to which phenotypic variance across environments could be explained by G, E, and

G×E. We found that early developmental traits were most strongly influenced by E (green-up = 98%, emergence time = 73%), followed by mid-stage traits (flowering time = 68%), and, finally, mature traits (plant height = 16%, tiller count = 14%, biomass = 26%). In contrast, the contributions of G×E were higher for mid-stage and mature traits (**Fig. 1a**). This pattern suggests that environmental effects contribute most to phenotypic variation in early developmental stages, whereas genetic factors contribute more as plants mature.

Given the environmental and G×E effects on trait variation, we next asked whether gene expression profiles similarly reflect environment-specific responses across genotypes. To test this, we collected transcriptome data from each genotype at both field sites (**see Methods**). Transcriptomic similarity among genotypes was measured using expression correlation (eCor), calculated as the PCC of gene expression levels. Clustering patterns of eCor differed between environments, indicating that while genotypes from similar genetic backgrounds (i.e., subpopulations; **Fig. 1b**) tend to cluster together, they also exhibit environment-specific transcriptional responses (e.g., clusters a–d; **Fig. 1b**). To investigate this further, we asked whether expression differences across environments ($\text{Diff}_{\text{exp}} = \text{TX} - \text{MI}$) reflect genetic similarity. If gene expression responses are genotype-driven, we would expect genetically similar individuals to show similar expression changes and cluster by subpopulation. Conversely, if responses are environment-specific, expression differences should not follow subpopulation structure. We observed no clear clustering by subpopulation in expression plasticity (**Fig. 1c**), indicating that across-environment transcriptional responses are primarily driven by environment rather than ancestry.

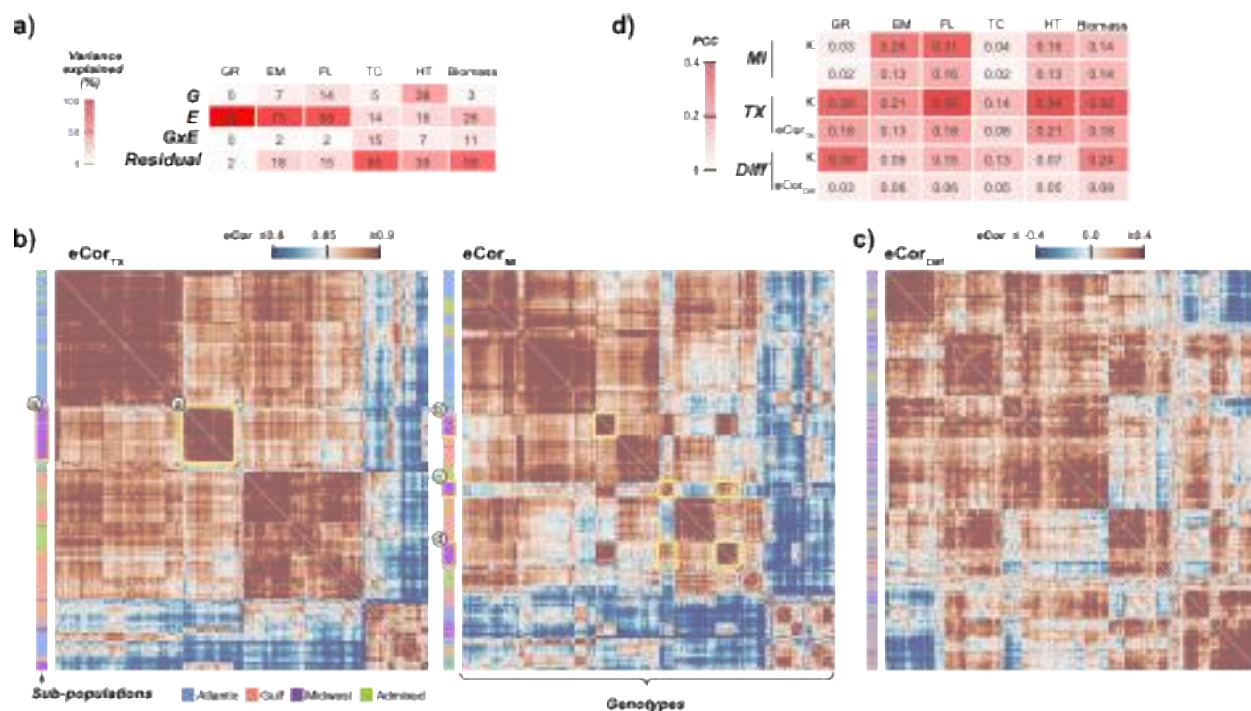


Figure 1. Relationships among phenotypes, gene expression, and genomic similarity across environments.

(a) Proportion of phenotypic variance explained by genotype (G), environment (E), and genotype-by-environment interaction (G×E) for six traits: green-up (GR), emergence (EM), flowering time (FLT), tiller count (TC), panicle height (HT), and log-transformed biomass. Pairwise expression correlation (eCor) among all genotypes based on transcriptomic data from **(b)** TX and MI (yellow rectangles highlight subpopulation-specific genotype clusters that shift across environments) and **(c)** gene expression differences between TX and MI (Diff), with heatmap colors ranging from blue (low) to white (intermediate) to brown (high) indicating the degree of correlation. **(d)** Pearson correlation coefficients (PCCs) between genomic relationship (K) or eCor based on TX, MI, or Diff transcript levels and phenotypic values in TX or MI or the difference in values between TX and MI (Diff). Heatmap colors ranging from white (zero) to red (positive) indicate PCC. Note that K calculated was using all genotypes, and the same value was used for all PCC calculations.

Having observed that gene expression within environments tends to cluster genotypes by genetic background, but transcriptional responses across environments do not, we next asked whether genetic similarity correlates with transcriptomic similarity. We found that genomic relatedness (kinship) derived from SNP data (**see Methods**) was significantly correlated with eCor within both TX ($r = 0.38$, $p < 2.2 \times 10^{-16}$) and MI ($r = 0.37$, $p < 2.2 \times 10^{-16}$), but was not correlated with Diff_{exp} ($r = 0.07$), consistent with clustering patterns observed in **Fig. 1b–c**. These results reinforce the observation that transcriptional responses across environments are decoupled from genetic similarity and instead largely shaped by environment. Because genotypes with similar genetic information are expected to exhibit similar trait values, we next assessed the relationships between kinship, eCor, and phenotypic traits within each environment. Overall, both kinship and eCor showed stronger correlations with trait values in TX than in MI (**Fig. 1d**), consistent with the higher proportion of trait variation explained by genetic effects (heritability) in TX for all traits except tiller count (**see next section**). Next, we tested whether variation in trait plasticity ($\text{Diff}_{\text{trait}} = \text{TX} - \text{MI}$) could be explained by either kinship or Diff_{exp} . Kinship partially explained variation in $\text{Diff}_{\text{trait}}$ for some traits, notably biomass ($r = 0.24$) and green-up ($r = 0.30$), while Diff_{exp} showed low correlation with trait plasticity across all traits ($r \leq 0.08$; **Fig. 1d**). These results demonstrate that genetic variants and gene expression contribute differently to trait variation and plasticity, reflecting distinct biological and environmental influences.

Omics data enable the prediction of traits in single environments and trait plasticity

To evaluate the extent to which transcriptomic and genetic variant (i.e., SNP) data predict switchgrass traits across environments, we built genomic prediction models using transcriptomic, SNPs, or both as input features (**Fig. 2**). SNPs remain constant for each genotype across environments, whereas transcriptomic data capture environmentally responsive changes in gene expression. Because trait variation involves both linear and nonlinear relationships among genes and traits¹⁵, we compared a linear (Bayesian Ridge Regression, BRR) with a nonlinear model (Extreme Gradient Boosting, XGBoost).

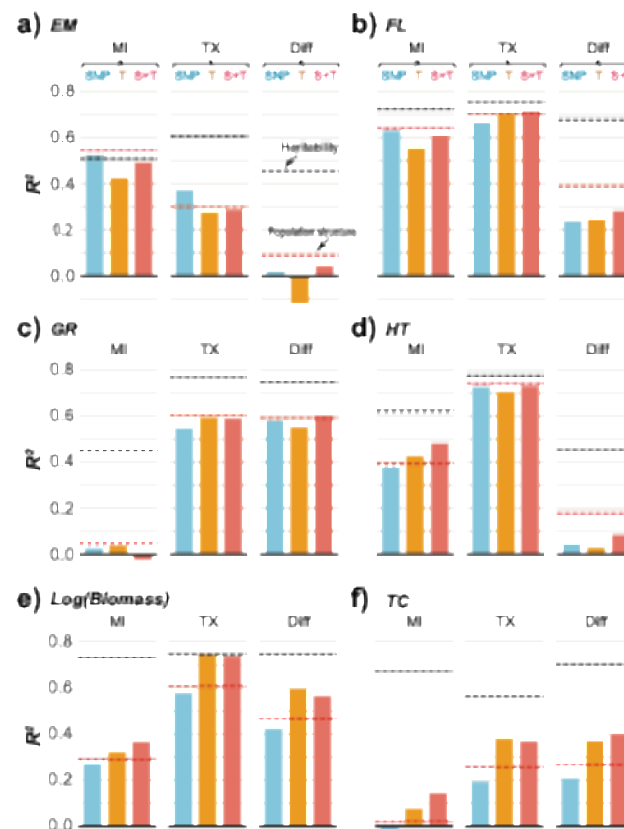


Figure 2. Model performance across locations and traits.

(a-f) XGBoost model performance (measured as R^2 using a testing set that has not been subjected to the training process) for six traits using genetic variants (G), transcriptomic variants (T), and their combination (G+T). Model performance based solely on population structure (first five principal components based on SNPs) is indicated by red dotted lines. Heritability estimates (black dotted lines, 0–1 scale) are shown for reference and are not equivalent to R^2 .

Model performance varied by trait but was generally consistent across algorithms and data types (**Supplementary table S1**). XGBoost models, except for biomass and tiller count models, did not outperform predictions based on the first five principal components of SNP data, which were used to approximate population structure (**Fig. 2**). Although we found that kinship correlated more strongly with phenotypic variation than eCor (**Fig. 1d**), both transcriptomic- and SNP-based models achieved comparable predictive performance. As expected, traits with higher correlations with genetic similarity (i.e., kinship or eCor; **Fig. 1d**) were predicted with higher accuracy (e.g., flowering time and biomass; **Fig. 2**). Notably, transcriptomic-based models of biomass consistently outperformed both SNP- and population structure-based models across sites, approaching the accuracy of heritability estimates in TX (**Fig. 2e**).

Because E and G×E effects together accounted for 20–98% of phenotypic variance across environments (**Fig. 1a**), we hypothesized that transcriptomic plasticity could serve as a predictor of trait plasticity (i.e., differences in trait values across environments). To test this, we trained models using SNP data, which remain constant across environments, and transcriptomic data, where plasticity was estimated as the difference in expression between TX and MI. We found that Diff_{exp} was more predictive

than SNPs for trait plasticity in biomass, flowering time, and tiller count, whereas SNPs were more predictive for green-up. For emergence and panicle height, however, both data types yielded low predictive performance ($R^2 < 0.1$), suggesting that the dataset lacked features capturing the plasticity of these traits. We further hypothesized that trait plasticity would be more predictable for traits with low phenotypic correlation between environments. Supporting this, plasticity SNP- and transcriptomic-based models generally outperformed single-environment models for traits with low cross-environment correlation, such as green-up (**Fig. 2c**, $r = -0.13$, $p < 8.3 \times 10^{-3}$) and tiller count (**Fig. 2f**, $r = -0.08$, $p = 0.1$) (**Supplementary Fig. S1**). In contrast, traits with high cross-environment correlation, such as flowering time and panicle height (**Fig. 2b,d**, $r > 0.68$, both $p < 2.2 \times 10^{-16}$), were less accurately predicted using difference-based models compared to single-environment models. Together, these results show that transcriptomic and SNP-based models capture complementary aspects of trait variation, including both within-environment effects and environmentally driven plasticity.

Transcript features enhance recovery of known and novel genes underlying trait variation and plasticity

Because SNP- and transcriptomic-based models yielded comparable predictive accuracy, we tested whether integrating both data types would improve performance. Multi-omic models that included both transcriptomic and SNP features outperformed either model alone in only 15% of cases, with modest gains ($R^2 \leq 0.07$); the largest improvement was observed for tiller count in MI (**Fig. 2f**). To investigate why combining SNP and transcriptomic data provides only limited benefit, we identified the genes associated with the SNPs and transcripts used by each model for prediction, defined as

those assigned a model importance score greater than zero (**Supplementary Table 2**). We focused on flowering time and biomass as examples because of their agronomic importance and found the highest gene overlap between SNP and transcriptomic models in TX for flowering time (60%), followed by the biomass plasticity model (45%). All other overlaps were $\leq 11\%$, indicating that models predicting the two data types largely rely on distinct sets of genes. Moreover, shared genes had uncorrelated SHAP values (**Supplementary Fig. S2**), consistent with prior findings in maize¹⁶, suggesting that SNP and transcriptomic models capture complementary information. To evaluate the relative contribution of each data type, we examined feature importance scores from multi-omic models. Transcriptomic features were consistently enriched among predictors with non-zero SHAP values (odds ratios = 5.32–14.8; all $p < 2.2 \times 10^{-16}$), indicating that transcriptomic data tend to be more informative than SNPs for trait prediction. This finding is consistent with prior results in maize¹⁶ and highlights the value of transcriptomic data in identifying functionally relevant genes.

To further evaluate the biological relevance of features, we asked whether transcriptomic or SNP data more effectively identify known causal genes. We assessed whether SNP- and transcriptomic-based models assigned non-zero feature importance to benchmark genes, defined as homologs of 1,552 flowering-related genes and 95 biomass-related genes from *Arabidopsis thaliana* and rice (*Oryza sativa*) (**Supplementary Table S3**), as well as 1,245 switchgrass biomass QTL genes from previous studies^{6,17,18} (**Supplementary Table S3; see Methods**). Across feature importance metrics (SHAP and BRR coefficients), transcriptomic models identified more benchmark genes with non-zero importance scores than SNP-based models. For

flowering time, 20% of benchmark genes (310 genes) were prioritized by transcriptomic models compared to 8% (129 genes) by SNP models (odds ratio = 2.3, $p=9.41 \times 10^{-15}$). A similar pattern was observed for biomass, with 18% (17 genes) vs. 4% (4 genes) prioritized, respectively (odds ratio = 4.4, $p=0.004$; **Supplementary Table S2**). These results suggest that transcriptomic models more effectively identify causative genes, potentially reflecting greater phenotypic influence of gene expression variation compared to SNP variation.

Having found that transcriptomic models incorporate more benchmark genes for prediction, we next asked which feature importance metric, SHAP or BRR coefficients, more effectively captures these genes in transcriptomic models. Because of the limited number of benchmark genes for biomass ($n = 17$), we focused this comparison on flowering time. SHAP consistently identified more flowering-related benchmark genes (130 for TX, 105 for MI, and 113 for Diff) than BRR coefficients (77, 54, and 62, respectively), indicating that SHAP provides a more informative measure of feature importance for identifying causal genes. We then evaluated whether benchmark genes were among the top predictors, as would be expected for causal loci. Using the top 5% of features ranked by importance, we observed significant enrichment for benchmark genes only in the TX flowering time model (odds ratio = 3.3, $p=1.73 \times 10^{-5}$). However, non-benchmark genes consistently received higher importance scores than benchmark genes across all models and metrics ($p<2.22 \times 10^{-16}$; **Supplementary Fig. S3**). These results suggest that while benchmark genes contribute to prediction, many genes contributing to variation in flowering time and biomass remain uncharacterized.

279 To identify genes driving model predictions, we examined the most influential genes in
 280 the transcriptomic-based models based on SHAP values. For flowering time, 12 genes
 281 were consistently ranked in the top 1% across both MI and TX (**Fig. 3a,c**), indicating
 282 that they represent a core set of genes that control flowering time across environments.
 283 This set included benchmark homologs of *FT* genes (*Pavir.4KG047800*,
 284 *Pavir.3KG349500*); genes in this family have been shown to promote flowering in
 285 multiple species, including *Arabidopsis* and switchgrass¹⁹. The core set also included a
 286 MADS-box gene (*Pavir.2KG001200*), and a non-benchmark bHLH transcription factor
 287 (*Pavir.1NG011300*) homologous to maize and sorghum genes involved in circadian
 288 regulation. Despite this shared core set, distinct top predictors emerged in each
 289 environment. For instance, the top predictor in TX was a kinesin motor protein
 290 (*Pavir.6NG016531*), while in MI, it was a *bZIP* transcription factor (*Pavir.9NG848400*);
 291 both are non-benchmark genes from families previously associated with flowering and
 292 plant development (Chen et al., 2024; Collani et al., 2019). In the flowering time
 293 plasticity model, *Pavir.4KG047800* (*FT*) was the top predictor (**Fig. 3a,d–e**), consistent
 294 with its role in environmental responsiveness in *Arabidopsis*²⁰. The second-ranked
 295 gene, *Pavir.9NG746100*, encodes an AP2/ERF transcription factor; members of this
 296 family are known to mediate responses to cold^{21,22} and regulate flowering under variable
 297 photoperiods²³.

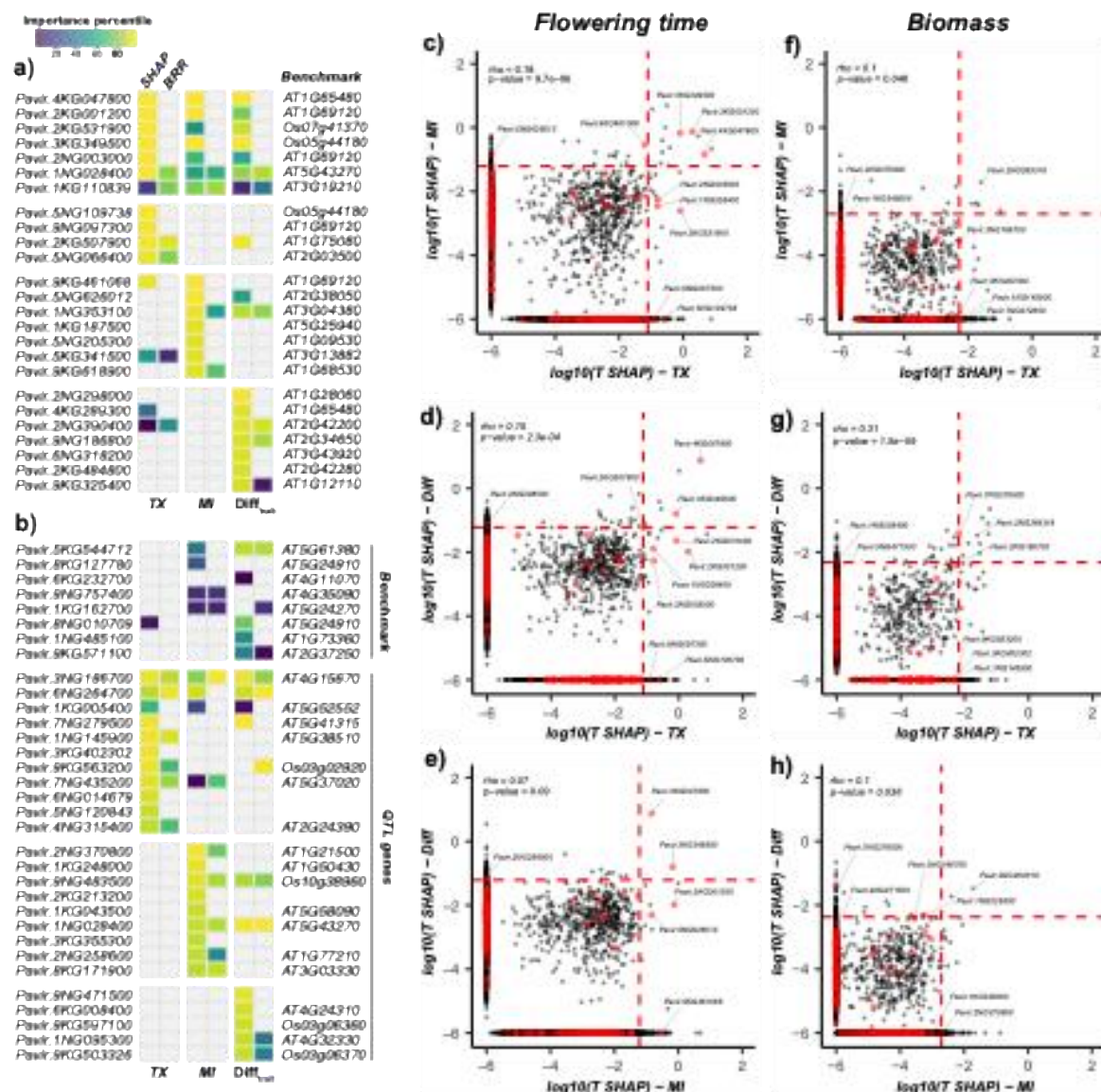


Figure 3. Correlation of transcript importance across MI, TX, and plasticity models for flowering time and biomass

(a-b) Feature importance scores for benchmark genes for flowering time (FLT) **(a)** and benchmark genes plus QTL candidate genes for biomass **(b)**. Feature importance was assessed using SHAP values and Bayesian ridge regression (BRR) for two individual environments (TX and MI) and their difference (Diff). Heatmap intensity represents the percentile rank of feature importance within each model, with yellow indicating highest

importance. **(c-h)** Comparison of SHAP values between the TX, MI, and Diff models for FL **(c-e)** and biomass **(f-h)**. SHAP values $< 10^{-6}$ were set to 10^{-6} for visualization. Red dashed lines indicate the 99th percentile of SHAP values. Red features represent benchmark genes **(c-e)** and QTL candidate genes **(f-h)**. Correlations are shown as Spearman rank coefficients (ρ).

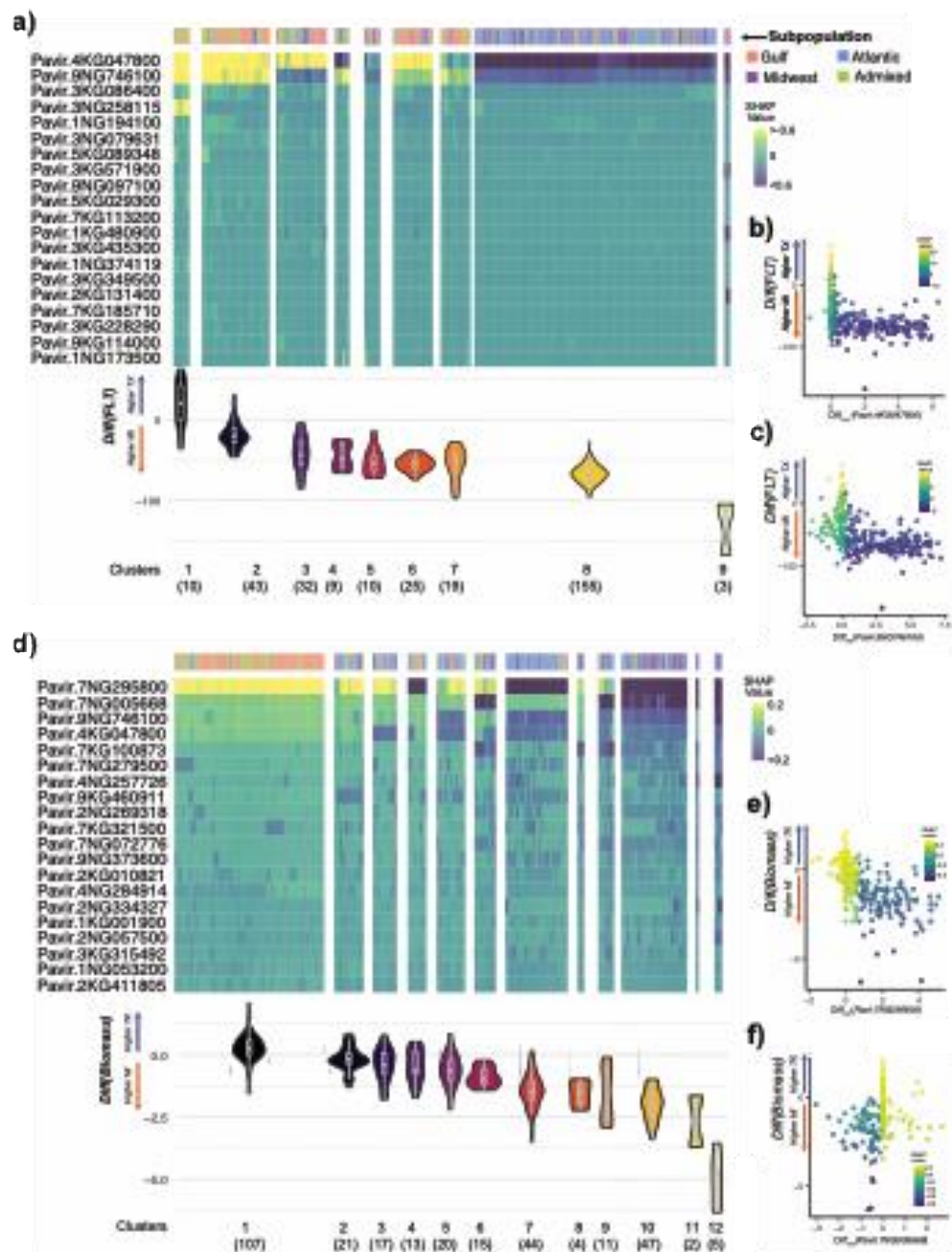
For biomass, five genes were among the top 1% of important predictors in both MI and TX, four of which encode uncharacterized proteins. The only annotated gene within this top 1% was *Pavir.2NG269318*, a meso-2,6-diaminoheptanedioate carboxy-lyase involved in lysine biosynthesis **(Fig. 3f)**. In *Arabidopsis*, lysine biosynthesis mutants show reduced photosynthesis and growth (Cavalcanti et al., 2018), supporting a functional link to biomass. This gene also ranked highly in the biomass plasticity model **(Fig. 3g-h)**. Environment-specific predictors included a kinesin motor protein (*Pavir.6NG016531*) in TX and a glycosyltransferase (*Pavir.1KG124131*) in MI; overexpression of glycosyltransferases has been linked to biomass increases in *Arabidopsis* (Zhao et al., 2025). The top predictor in the biomass plasticity model was an AP2/ERF transcription factor (*Pavir.7NG295800*), previously associated with biomass in switchgrass (Lovell et al., 2021). As in flowering time, AP2/ERFs emerged as regulators of G×E responses, consistent with their roles in mediating plant responses to climatic and photoperiod variation^{21–23}. Taken together, these results demonstrate that transcriptomic models capture both shared and environment-specific biological signals, effectively identifying benchmark genes and uncovering novel, uncharacterized candidates that may underlie natural variation in traits and their plasticity.

Gene effects are genotype and environment dependent

Building on our finding that models recover both benchmark and novel candidate genes, we next asked whether gene effects vary among genotypes across environments. To quantify genotype-specific contributions of individual genes, we computed local SHAP values for each genotype, where positive SHAP values indicate that higher feature values (e.g., gene expression) are associated with higher trait values, and negative values indicate the opposite. We focused on the top 20 genes predictive of FLT and biomass, and used SHAP profiles to cluster genotypes based on shared patterns of gene importance. In each data set, these clusters showed distinct trait distributions (**Supplementary Fig. S4, S5**), demonstrating that SHAP profiles reflect transcriptomic influences on trait variation and capture genotype-level differences in how genes contribute to trait expression.

To investigate G×E interactions, we analyzed transcriptome-based models predicting trait value differences across environments. Clustering genotypes based on their local SHAP profiles revealed nine distinct groups with varying FLT plasticity (**Fig. 4a**). In these models, positive SHAP values indicate the contribution of predictors to delayed FLT in TX relative to MI, while negative values indicate earlier FLT in TX. Clustering was primarily driven by *Pavir.4KG047800* (FT homolog) and *Pavir.9NG746100* (AP2/ERF transcription factor). Genotypes in clusters 1 and 2 exhibited positive SHAP values, corresponding to delayed FLT in TX. To understand this variation, we examined gene expression differences across environments. Genotypes with higher FT and AP2/ERF expression in TX tended to flower earlier in TX compared to MI, while genotypes with stable expression tended to flower later in TX (**Fig. 4b–c**). These results

351 suggest that expression plasticity of FT and AP2/ERF contributes to adaptive shifts in
 352 FLT. Consistent with this, FT and AP2/ERF genes are known regulators of flowering
 353 and environmental responsiveness across species^{2,24–27}.



354

Figure 4. Genotype-dependent effects of top transcriptomic predictors of trait plasticity

(a, d) SHAP values for the top 20 genes (y-axis) across genotypes (x-axis) for flowering time **(a)** and biomass **(d)** plasticity models. For visualization, SHAP values were clipped to ± 0.6 for flowering time and ± 0.2 for biomass **(b–c, e–f)** Relationship between gene expression plasticity (difference in expression level [TPM] between TX and MI; x-axis) and trait plasticity (difference in trait value between TX and MI; y-axis) for the top two predictors of FLT **(b–c)** and biomass **(e–f)**. Colors indicate the SHAP values.

In the biomass model, SHAP values grouped genotypes into 12 distinct clusters (Fig. 4d). The top predictors included two AP2/ERF transcription factors (Pavir.7NG295800 and Pavir.9NG746100, ranked 1 and 3), a gene of unknown function (Pavir.7NG005668, rank 2), and FT (Pavir.4KG047800, rank 4). Notably, both Pavir.9NG746100 and FT were also among the top predictors for FLT, reinforcing their roles in regulating G×E responses across traits²⁶. Clustering revealed distinct biomass patterns. Cluster 1 included genotypes with higher biomass in TX compared to MI, genotypes in clusters 2–5 showed similar biomass across environments, and those in clusters 6–12 had higher biomass in MI compared to TX (**Fig. 4d**). The top predictors exhibited opposing SHAP value directions across clusters (e.g., clusters 3–7, 9, 11), indicating decoupled gene effects. For instance, genotypes with higher expression of *Pavir.7NG295800* in TX had negative SHAP values (i.e., higher biomass in MI; **Fig. 4e**), while *Pavir.7NG005668* showed the opposite pattern (**Fig. 4f**). Similar genotype-specific contrasting SHAP values were observed for the top 20 predictors in single-environment models (**Supplementary Fig. S4, S5**), highlighting the genotype dependent effect of T features. Together, these findings demonstrate that gene effects on FLT and biomass

are both genotype and environment dependent and that variation in gene expression contributes to trait plasticity across environments.

Gene-by-gene interactions reveal environment-specific regulatory architecture

In previous sections, we showed that gene effects on traits are both genotype- and environment-dependent. Because each genotype carries a unique combination of alleles, we hypothesized that variation in gene composition among genotypes could contribute to differences in gene–gene interactions, ultimately influencing trait expression. To evaluate these interactions and how they vary across environments, we used SHAP interaction scores²⁸, which quantify pairwise feature interactions (i.e., G×G) as a proxy for epistasis. We first asked whether the number and magnitude of interactions were consistent across models. Genes important for FLT showed both stronger and more interactions than those for biomass (**Fig. 5a–b**). This likely reflects the underlying genetic architecture of FLT, which is regulated by central hubs with large effect sizes^{29,30}, in contrast to the more diffuse, polygenic control of biomass. Such large-effect loci may enhance the detectability of interaction signals.

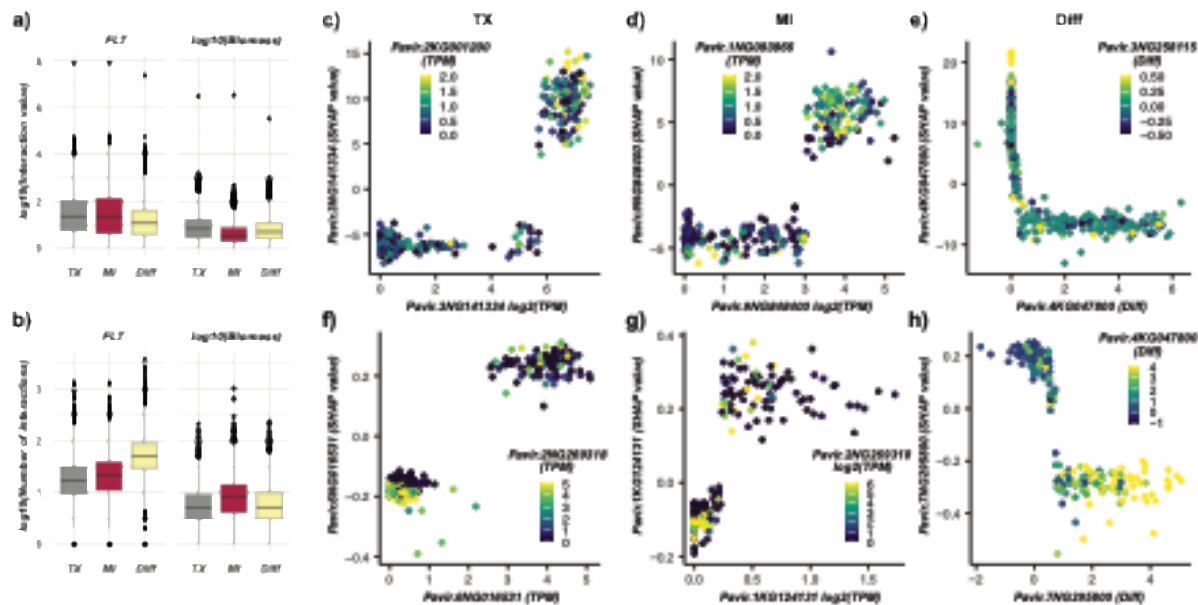


Figure 5. Gene-gene interaction profiles associated with flowering time and biomass.

(a) Distribution of SHAP interaction values for the flowering time (FLT; left), and biomass (right) models. **(b)** Number of gene-gene interactions per gene for FLT (left) and biomass (right). **(c-d)** examples of top interactions for FLT **(c)**, and biomass **(d)**.

Genes with strong interactions can modulate the effects of other genes and contribute to non-additive genetic variance. To identify hub genes, we ranked the top 100 genes by the number of SHAP interactions. In FLT models, two FT homologs (*Pavir.4KG047800*, *Pavir.3KG349500*) and two MADS-box genes consistently ranked among the top interactors in both MI and TX, each with over 169 interactions (**Supplementary Table S5**). Additionally, two non-benchmark genes ranked first in MI (*Pavir.9NG848400*, a bZIP transcription factor) and TX (*Pavir.3NG141334*, an FKBP). Both gene families have previously been implicated in flowering regulation^{31,32}, supporting their potential roles as FLT hub genes. In the plasticity model,

Pavir.4KG047800 (FT) had the highest interaction count (3,670), reinforcing the hypothesis that it plays a central role in mediating G×E responses. Other top benchmark genes included BES/BZR, AGC kinase, and bHLH transcription factors, all known regulators of environmental responses^{33–35}. Top non-benchmark interactors included a cytochrome P450 and an AP2-like transcription factor, members of gene families linked to stress adaptation and plasticity^{36,37}. Together, these results demonstrate that T models capture both canonical flowering regulators and novel candidate co-regulators of FLT and its plasticity.

In contrast, for biomass models, no benchmark genes were among the top 100 SHAP interactors; instead, several non-benchmark genes with potential roles in biomass-related traits were identified (**Supplementary Table S6**). For example, in TX, *Pavir.5NG530000* (*ULTRAPETALA* homolog) ranked second (166), and in MI, the top interactor was *Pavir.1KG124131* (*Glycosyltransferase*; 1005). *ULTRAPETALA* genes have been linked to meristem size and plant architecture³⁸, while glycosyltransferases have been associated with biomass accumulation³⁹. In the plasticity model, the top interactor was *Pavir.7NG295800* (AP2/ERF; 233). Notably, the FT homolog *Pavir.4KG047800* ranked 11th (105), highlighting the potential role of the vegetative-to-reproductive transition in biomass accumulation and its contribution to plasticity.

Examining interactions in detail, we found that *Pavir.4KG047800* (FT homolog) and *Pavir.2KG001200* (*MADS-box AP1* homolog) were co-expressed in both MI and TX, where high expression correlated with earlier flowering (**Supplementary Fig. S6a–b**), consistent with FT’s activation of *AP1*⁴⁰. In the plasticity model, interactions between two FT homologs (*Pavir.4KG047800* and *Pavir.3KG349500*) showed environment-

dependent *effects*; genotypes that failed to upregulate FT-like genes in Texas flowered later, underscoring FT's central role in flowering plasticity (**Supplementary Fig. S6c**). We also identified a non-benchmark gene, *Pavir.3NG141334* (an FKBP-like gene), which interacts with *AP1* and is associated with delayed flowering in TX when both genes show high expression (**Fig. 5c**), paralleling *FKBP12–CONSTANS* interactions in *Arabidopsis*³². In MI, we identified an interaction between a *bZIP* (*Pavir.9NG848400*) and a glucosyltransferase (*Pavir.1NG083866*) (**Fig. 5d**). Members of both genes families have established roles in *FT* activation complexes and hormonal regulation of flowering^{40,41}. Another glucosyltransferase (*Pavir.3NG258115*) interacted with *Pavir.4KG047800* (*FT*) in the plasticity model, where genotypes with stable *FT* and high *Pavir.3NG258115* expression flowered later in TX (**Fig. 5e**).

SHAP interaction values revealed that interactions between *Pavir.2NG269318* (carboxylase) and *Pavir.6NG016531* (kinesin motor protein) in Texas, and between *Pavir.2NG269318* and *Pavir.1KG124131* (glycosyltransferase) in Michigan, contributed to model predictions of biomass (**Fig. 5f–g**). In both cases, high *Pavir.2NG269318* expression was associated with negative SHAP values for *Pavir.6NG016531* and *Pavir.1KG124131*, indicating that under high *Pavir.2NG269318* expression, these growth-related genes contribute negatively to biomass. One potential explanation for these interactions is that high *Pavir.2NG269318* expression alters metabolic flux in a way that limits the effectiveness of cell expansion, mediated by kinesin motor proteins⁴², and cell-wall biosynthesis, driven by glycosyltransferases⁴³. In the plasticity model, *Pavir.7NG295800* (AP2/ERF) interacted with *Pavir.4KG047800* (*FT*), where higher expression in TX relative to MI for both genes corresponded to reduced biomass in TX

compared to MI. This pattern may indicate that elevated expression of *FT* (a flowering regulator) and an AP2/ERF transcription factor (a regulator of growth and development)²⁶ shortens the vegetative phase in TX, resulting in lower biomass accumulation. Together, these results suggest that local interpretation can recover known flowering regulators and highlight novel, environment-specific gene relationships, offering a framework for generating hypotheses about the molecular basis of flowering, biomass accumulation, and G×E plasticity.

Conclusion

By integrating genomic and transcriptomic data with interpretable machine learning, we show that (i) information from models based on transcripts contributing environment-sensitive signals doesn't overlap with that captured by models using genetic variance; (ii) gene effects are strongly genotype- and environment-dependent; and (iii) local interpretation reveals biologically meaningful, environment-specific gene–gene interactions. *FT* homologs consistently emerged as central hubs for flowering time and, to a lesser extent, biomass, highlighting the pleiotropic role of flowering regulators in connecting reproductive transitions with vegetative growth. These findings translate complex model predictions into mechanistic hypotheses about causal genes and interactions underlying trait variation, both within single environments and across environments. Together, our results prioritize candidate regulators for experimental validation and provide a general framework for dissecting the genetic basis of traits and their plasticity in plants.

Materials and Methods

Phenotypic data

The phenotypic data used in this study were previously described⁶. Briefly, the switchgrass diversity panel was established in 2018, with spaced plants arranged in a honeycomb pattern across 10 sites in the US. The locations in Michigan and Texas were chosen for the current study because of their contrasting environmental conditions and latitudes, which have been observed to influence important trait variation in switchgrass. In Michigan, the diversity panel was grown at the Michigan State University Kellogg Biological Station (KBSM) (latitude: 42.419618, longitude: - 85.371266, elevation: 289 meters). In Texas, plants were grown at the University of Texas J.J. Pickle Research Campus (latitude: 30.383979, longitude: -97.729383, elevation: 235 meters). Genotypes were replicated across field sites and haphazardly randomized within each of the two field site locations. In total, six traits were measured in 2021 for a subset of 426 genotypes: green-up (GR), emergence (EM), flowering time (FLT), tiller count (TC), height of the panicle (HT), and biomass. GR was measured as the day of the year when plants began to green up. EM was measured as the day of the year when panicles began to emerge. FL was recorded as the day of the year when flowering started. TC and HT were measured at the end of the season, with height recorded as the mean panicle apex in centimeters. For biomass, whole plants were harvested and weighed (recorded in grams), and dry weight was estimated for each sample based on a ~500 g subsample of fresh material. The estimated dry biomasses were log-transformed to normalize the data. Phenotypic data and correlations across locations are presented in **Supplementary Figure 1**. To partition phenotypic variance

into genetic (G), environmental (E), and genotype-by-environment interaction (G×E) components, we fit a linear mixed model using the sommer R package⁴⁴. The model included E and G×E as random effects, and G was modeled using a genomic relationship matrix (kinship). The model was specified as: $y \sim 1 + (1 | E) + (1 | G) + (1 | G:E)$, where genotype was structured using the kinship matrix. Variance components were extracted from the fitted model, and the proportion of total variance explained by each component was calculated.

Whole genome data and genetic variants

Whole-genome sequencing data from the switchgrass diversity panel, comprising 726 genotypes, were downloaded from the NCBI SRA under BioProject PRJNA622568. The genotypes were sequenced using Illumina HiSeq X10 and Illumina NovaSeq 6000 paired-end sequencing (2 × 150 bp) at the HudsonAlpha Institute for Biotechnology and the Joint Genome Institute⁶. Genetic variant detection followed the GATK Best Practices (Broad Institute). Briefly, raw reads were first trimmed as paired-end reads using fastp with default parameters⁴⁵. Clean reads were then mapped to the AP13 HAP1 v6.1 genome (used with permission) using BWA-MEM⁴⁶ with default parameters. The resulting BAM files were sorted using samtools⁴⁷. Next, duplicate reads were marked using picard toolkit MarkDuplicates. Variant calling was performed using GATK HaplotypeCaller, generating a GVCF (genomic variant calling file) for each sample. The GATK GenomicsDBImport function was then used to merge GVCFs from all 732 samples into a GenomicsDB. Finally, GATK GenotypeGVCFs was used to identify genetic variants in each sample. The resulting VCF file was filtered using bcftools to retain the 426 genotypes from which phenotypic data were collected, as well as SNPs

with no missing data, a minor allele frequency (MAF) > 0.5, and linkage disequilibrium (LD) > 0.1. The final VCF file contained 220,263 SNPs.

Leaf collections and RNA extractions

Leaves of mature plants for each of the 426 genotypes were collected from the field in both in TX and MI. We timed collections at both sites to best approximate the same pre-flowering developmental stage. Leaf collections at the common garden located at J.J. Pickle Research Campus in Austin, TX were conducted on May 5 and May 6, 2021. Leaf collections from the common garden at the Kellogg Biological Station in Hickory Corners, MI were conducted on June 22nd and June 24th, 2021. Leaves at both locations were collected between 10 am and noon. For each collection, the bases of each leaf were placed into 15mL tubes and flash frozen in liquid nitrogen within less than a minute of removal from each plant. Tissue was then transported to -80 freezers at The University of Texas and Michigan State University on dry ice.

For RNA extraction, approximately 200 mg of leaf tissue was homogenized into a fine powder in 2.0-mL Premium Microcentrifuge Tubes (Fisher Scientific) using stainless steel beads on a Geno/Grinder 2000 (Spex SamplePrep). The tubes were kept in a pre-frozen microcentrifuge tube rack under liquid nitrogen to maintain the samples in a frozen state during homogenization. Total RNA was isolated using the Direct-zol™ RNA Miniprep Kit (Zymo Research) following the manufacturer's instructions. In brief, 1 mL of TRIzol™ Reagent (Invitrogen) was added immediately to the frozen powdered tissue, followed by vigorous vortexing to fully resuspend the homogenate. The lysate was centrifuged at 12,000 × g for 5 min at 4°C to remove insoluble debris, and 400 µL of the

clear supernatant was transferred to a new RNase-free microcentrifuge tube. An equal volume (400 μ L) of 100% ethanol was added to the supernatant and mixed thoroughly. The mixture was then loaded onto a Direct-zol™ spin column, and RNA purification and on-column DNase I treatment were performed according to the manufacturer's protocol. RNA was eluted by adding 40 μ L of DNase/RNase-free water directly to the column matrix and centrifuging. The integrity and concentration of the RNA preparations were checked initially using Nano-Drop (Nano-Drop Technologies) and then by BioAnalyzer (Agilent Technologies).

RNA-seq data normalization

RNA-seq was performed using NovaSeq 6000 paired-end sequencing (2 x 150 bp) at the Joint Genome Institute. Raw fastq file reads were filtered and trimmed using the JGI QC pipeline. Raw reads were evaluated for artifact sequence by kmer matching (kmer=25), allowing 1 mismatch, and detected artifacts were trimmed from the 3' end of the reads using BBduk, a tool included in the BBDuk package (<https://sourceforge.net/projects/bbmap/>). RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, reads under 50 bases in length were removed. Filtered reads from each library were aligned to the AP13 HAP1 v6.1 genome using HISAT2 version 2.2.0⁴⁸. Strand-specific coverage bigWig files (fwd and rev) were generated using deepTools v3.1⁴⁹. featureCounts⁵⁰ was used to generate the raw gene counts file using AP13 HAP1 v6.1 gff3 annotations. Only primary hits assigned to the reverse strand were included in the raw gene counts. For downstream machine learning analysis, raw read counts for each gene were divided by the gene's length (in kilobases)

to calculate Reads Per Kilobase (RPK). Then, RPK values for all genes within a sample were summed, and each RPK value was divided by this sum to obtain Transcripts Per Million (TPM) values. To remove genes with zero or nearly zero variance (>95% of samples sharing the same transcript level), we used the nearZeroVar function from the R caret package. After filtering, TPM counts for 49,985 genes were retained in the final dataset. Before further analysis, TPM values were transformed ($\log_{10}(\text{TPM} + 1)$).

Correlations between genetic relationships, gene expression, and phenotypic variation across environments

The genomic relationship matrix (G) was estimated using centered and standardized SNP information, following the formula $G = ZZ'/p$ ⁵¹, where p represents the number of SNPs and Z is the matrix of centered and standardized SNPs. For gene expression, Pearson correlations of TPM values between genotypes were computed using the cor.test function in R, generating an expression correlation (eCor) matrix for each environment. For phenotypic data, the Euclidean distance between genotypes was calculated for each trait. The correlation between the G, eCor, and phenotypic distance was assessed using Pearson correlation. Additionally, genomic heritability (h_g^2), defined as the proportion of variance in a trait that can be explained by a linear regression on markers⁵², was estimated for all traits.

Predictive modeling and feature importance

Because of the well-known relationship between subpopulations and phenotypic variation in switchgrass, a baseline for the prediction models was established using the first five principal components (PCs) derived from SNPs to predict phenotypic values.

The first five PCs were chosen as a proxy for population structure because, beyond the fifth PC, the proportion of variance explained is lower than 1% (Supplementary Figure 2). Genotypes for the training and test sets were selected using stratified sampling with ten equally spaced quantile bins (10% intervals from 0 to 1), ensuring proportional sampling of individuals in both sets from each quantile. To maintain consistency across traits and environments, this stratification approach was applied to log-transformed biomass in TX, ensuring that the same individuals were used in the training and test sets for all traits and environments.

For each trait, predictions were performed using a linear model, Bayesian Linear Regression (BRR), and a non-linear model, eXtreme Gradient Boosting (XGBoost). The BRR models were implemented in R using the BGLR package⁵³ with 12,000 iterations, where the first 2,000 iterations were discarded as burn-in. XGBoost was implemented in Python using the Scikit-Learn package⁵⁴. To optimize the hyperparameters of the XGBoost models, a Bayesian optimization using the Tree-structured Parzen Estimator (TPE) algorithm was implemented using the HyperOpt package⁵⁵. The hyperparameter search space was defined as follows: the learning rate was sampled from a uniform distribution between 0.001 and 0.3; the maximum tree depth was chosen from discrete values (3, 5, 10); the subsample ratio was sampled from a uniform distribution between 0.8 and 0.9; feature subsampling was sampled from a uniform distribution between 0.5 and 1.0; and the number of estimators was chosen from discrete values (50, 100, 150, 200). The optimization process was assessed using R^2 through 5-fold cross-validation. Note that Figure 2 includes only model performance on the testing dataset. Cross-validation was performed within the training set for hyperparameter tuning. Once the

optimal parameters were identified, the model was retrained on the full training set and used to predict the testing dataset. Cross-validation R^2 values are provided in Supplementary Table 1. The search process was executed for 100 iterations, after which the best-performing hyperparameters were selected. A final model was then built using all genotypes in the training set and applied to the test set to evaluate performance using R^2 .

To evaluate the global contribution of features to model predictions, feature importance values were obtained from both the BRR and XGBoost models. For BRR, feature importance was measured as the estimated posterior mean of feature effects, calculated using the R package BGLR⁵³, with the absolute values of these effects used for interpretation. For XGBoost, feature importance was assessed using gain-based importance, extracted from the fitted models using the “feature_importances_” function from the Scikit-Learn package⁵⁴. To reduce variability in feature importance estimates caused by stochastic elements in the boosting process, such as random subsampling and feature selection, the importance scores were averaged across 10 model runs using the same set of optimized hyperparameters and training dataset. To assess the local contribution of features to individual predictions (i.e., genotypes), SHapley Additive Explanations (SHAP) values⁵⁶ were obtained from XGBoost. SHAP values were computed in each of the 10 runs using the “explainer_model” function from the SHAP package⁵⁶, and were then averaged across runs.

Feature selection and feature interactions

Feature importance from XGBoost models was used for feature selection. The importance scores were averaged across 10 model runs, and only features with non-

zero importance were retained for model prediction. These features were then ranked based on their importance scores from the initial model, which included all features. Subsets containing the top 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 8, 6, 4, 2, and 1 percent of features with non-zero importance were selected. New models were trained using each feature subset on the training dataset, following the same process for hyperparameter optimization and cross-validation as described for the initial models (**Supplementary table 1**). The best feature subset was selected based on the R^2 values of the fitted models. Models using the best feature subset were used to estimate the interaction among features using the “shap_interaction_values” function from the SHAP package⁵⁶.

Identification of candidate genes for biomass

For flowering time, we downloaded 687 and 157 benchmark flowering-time genes (**Supplementary table 3**) from the TAIR and FLOR-ID databases, respectively; these genes are known to regulate flowering in *Arabidopsis*. We also included 237 switchgrass genes associated with flowering time and homologous to 130 *Arabidopsis* and 107 rice genes previously identified in the literature and databases as flowering-time genes^{57,58}, resulting in a total of 1,081 genes.

For biomass, we downloaded 63 benchmark genes from TAIR (**Supplementary table 3**). Because few experimentally validated biomass genes are available, we also compiled 1,896 candidate genes from previous studies: 32 from QTL analyses^{17,18} and 1,864 from GWAS analyses⁶. These candidates were originally identified using earlier versions of the switchgrass reference genome (v1.1, v3.1, v4.1, and v5.1). To map them to the most recent reference genome (AP13 v6.1, Hap1), we used OrthoFinder and

GENESPACE with default parameters, restricting homolog identification to syntenic orthologs to ensure reliable cross-version mapping (**Supplementary table 4**).

Data availability

The raw RNA sequencing data for all libraries are available under BioProject PRJNA1402996 at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1402996>

Code availability

Custom pipelines for SNP calling, kinship, expression correlation matrices, heritability, predictive modeling, and identification of syntenic orthologous genes are available at <https://github.com/pauloizquierdo/switchgrass-GxE-modeling.git>

Acknowledgments

We thank the field technicians and volunteers who maintained field sites and assisted with phenotype and leaf tissue collections, including L. Vormwald, M. Ryskamp, P. Wang, N. Emery, A. VanWallendael, T. Zambiasi, K. Segura Abá, T. Ranaweera, J. Yin and C. Malmstrom. We thank the Department of Energy Joint Genome Institute and collaborators for pre-publication access to the genome of *Panicum virgatum* AP13 HAP1 v6.1 for RNA-seq analysis. This work was supported by the U.S. Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research under Awards DE-SC0018409 and DE-SC00141156. Support was also provided by the National Science Foundation Long-Term Ecological Research Program (DEB-1832042) at Kellogg Biological Station. Work conducted by the DOE Joint Genome Institute (proposal 10.46936/10.25585/60000516), a DOE Office of Science User Facility, is

supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author Contributions

P.I., D.L., T.J., M.L., and S.H.S. designed research. P.I., D.L., T.J., X.W., Y.Y., C.D., A.L., J.B., and S.H.S. performed research. P.I., and S.H.H analyzed data; and P.I., D.L., M.L., and S.H.H wrote the paper. All authors read and approved the final manuscript.

Supplementary figures

Supplementary Figure 1. Pearson correlation (r) between traits across environments (x-axis: MI; y-axis: TX) for green-up (GR), emergence (EM), flowering time (FLT), tiller count (TC), panicle height (HT), and log-transformed biomass (measured in g).

Supplementary Figure 2. Density scatterplots of importance scores for SNPs (y-axis) and transcripts (x-axis) for SNP:T pairs associated with flowering time (FLT) (a–c) and biomass (d–f) in TX, MI, and difference models. Thresholds (yellow dotted lines) represent the 99th percentile of importance scores. Spearman's rank correlation (ρ) quantifies correspondence between SNP and transcript importance.

Supplementary Figure 3. Comparison of transcriptomic importance scores for benchmark and non-benchmark genes associated with flowering time and biomass. p-values are from two-sided Wilcoxon rank-sum tests.

Supplementary Figure 4. Genotype-dependent effects of top transcriptomic predictors of flowering time in TX (top), MI (middle), and difference models (bottom).

Supplementary Figure 5. Genotype-dependent effects of top transcriptomic predictors of biomass in TX (top), MI (middle), and difference models (bottom).

Supplementary Figure 6. Gene–gene interaction profiles for benchmark flowering time genes, including FT and MADS-box genes, in TX (a), MI (b), and two FT genes in the difference model (c).

Supplementary tables

Supplementary Table 1. Model training/testing performance and feature selection summary across datasets.

Supplementary Table 2. Feature importance scores and benchmark gene annotations

Supplementary Table 3. Switchgrass homologs of benchmark genes in Arabidopsis and rice for flowering time and biomass

Supplementary Table 4. Candidate biomass genes identified through QTL and GWAS in switchgrass

Supplementary Table 5. Top 100 interacting features identified through SHAP interaction values for flowering time

Supplementary Table 6. Top 100 interacting features identified through SHAP interaction values for biomass

References

1. Anderson, J. T., Lee, C.-R., Rushworth, C. A., Colautti, R. I. & Mitchell-Olds, T. Genetic trade-offs and conditional neutrality contribute to local adaptation. *Mol. Ecol.* **22**, 699–708 (2013).
2. Des Marais, D. L., Hernandez, K. M. & Juenger, T. E. Genotype-by-Environment Interaction and Plasticity: Exploring Genomic Responses of Plants to the Abiotic Environment. *Annu. Rev. Ecol. Evol. Syst.* **44**, 5–29 (2013).
3. Napier, J. D., Heckman, R. W. & Juenger, T. E. Gene-by-environment interactions in plants: Molecular mechanisms, environmental drivers, and adaptive plasticity. *Plant Cell* **35**, 109–124 (2023).
4. Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nat. Rev. Genet.* **14**, 807–820 (2013).
5. Casler, M. D., Vogel, K. P. & Harrison, M. Switchgrass Germplasm Resources. *Crop Sci.* **55**, 2463–2478 (2015).
6. Lovell, J. T. *et al.* Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* **590**, 438–444 (2021).
7. Lowry, D. B. *et al.* QTL × environment interactions underlie adaptive divergence in switchgrass across a large latitudinal gradient. *Proc. Natl. Acad. Sci.* **116**, 12933–12941 (2019).
8. Wang *et al.* Cropformer: An interpretable deep learning framework for crop genomic prediction. *Plant Commun.* **6**, (2025).
9. Yu *et al.* EXGEP: a framework for predicting genotype-by-environment interactions using ensembles of explainable machine-learning models. *Brief. Bioinform.* **26**, bbaf414 (2025).

10. Singhal, R. *et al.* Using supervised machine-learning approaches to understand abiotic stress tolerance and design resilient crops. *Philos. Trans. R. Soc. B Biol. Sci.* **380**, 20240252 (2025).
11. Lundberg, S. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
12. Johnsen, P. V., Riemer-Sørensen, S., DeWan, A. T., Cahill, M. E. & Langaas, M. A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values. *BMC Bioinformatics* **22**, 230 (2021).
13. Wang, P. *et al.* Prediction of plant complex traits via integration of multi-omics data. *Nat. Commun.* **15**, 6856 (2024).
14. Falconer, D. S. & Mackay, T. *Introduction To Quantitative Genetics 4th Edition*. (1996).
15. Mackay, T. F. C. Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions. *Nat. Rev. Genet.* **15**, 22–33 (2014).
16. Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G. & Shiu, S.-H. Transcriptome-Based Prediction of Complex Traits in Maize[OPEN]. *Plant Cell* **32**, 139–151 (2020).
17. Razar, R. M., Qi, P., Devos, K. M. & Missaoui, A. M. Genotyping-by-Sequencing and QTL Mapping of Biomass Yield in Two Switchgrass F1 Populations (Lowland x Coastal and Coastal x Upland). *Front. Plant Sci.* **13**, (2022).
18. Shrestha, S. L. *et al.* Mapping quantitative trait loci for biomass yield and yield-related traits in lowland switchgrass (*Panicum virgatum* L.) multiple populations. *G3 GenesGenomesGenetics* **13**, jkad061 (2023).
19. Niu, L. *et al.* Control of floral transition in the bioenergy crop switchgrass. *Plant Cell Environ.* **39**, 2158–2171 (2016).
20. Sasaki, E., Zhang, P., Atwell, S., Meng, D. & Nordborg, M. ‘Missing’ G x E Variation Controls Flowering Time in *Arabidopsis thaliana*. *PLOS Genet.* **11**, e1005597 (2015).

21. Gu, C. *et al.* Multiple regulatory roles of AP2/ERF transcription factor in angiosperm. *Bot. Stud.* **58**, 6 (2017).
22. Zhang, H. *et al.* Genome-wide identification and comprehensive analysis of the AP2/ERF gene family in *Prunus sibirica* under low-temperature stress. *BMC Plant Biol.* **24**, 883 (2024).
23. Shim, Y. *et al.* The AP2/ERF transcription factor LATE FLOWERING SEMI-DWARF suppresses long-day-dependent repression of flowering. *Plant Cell Environ.* **45**, 2446–2459 (2022).
24. Böhlenius, H. *et al.* CO/FT Regulatory Module Controls Timing of Flowering and Seasonal Growth Cessation in Trees. *Science* **312**, 1040–1043 (2006).
25. Li, H. *et al.* The AP2/ERF transcription factor TOE4b regulates photoperiodic flowering and grain yield per plant in soybean. *Plant Biotechnol. J.* **21**, 1682–1694 (2023).
26. Ma, Z., Hu, L. & Jiang, W. Understanding AP2/ERF Transcription Factor Responses and Tolerance to Various Abiotic Stresses in Plants: A Comprehensive Review. *Int. J. Mol. Sci.* **25**, 893 (2024).
27. Wigge, P. A. *et al.* Integration of Spatial and Temporal Information During Floral Induction in *Arabidopsis*. *Science* **309**, 1056–1059 (2005).
28. Zern, A., Broelemann, K. & Kasneci, G. Interventional SHAP Values and Interaction Values for Piecewise Linear Regression Trees. *Proc. AAAI Conf. Artif. Intell.* **37**, 11164–11173 (2023).
29. Chen, Yan, W., Fu, L.-Y. & Kaufmann, K. Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. *Nat. Commun.* **9**, 4534 (2018).
30. Madrid, E., Chandler, J. W. & Coupland, G. Gene regulatory networks controlled by FLOWERING LOCUS C that confer variation in seasonal flowering and life history. *J. Exp. Bot.* **72**, 4–14 (2021).

31. Collani, S., Neumann, M., Yant, L. & Schmid, M. FT Modulates Genome-Wide DNA-Binding of the bZIP Transcription Factor FD1[OPEN]. *Plant Physiol.* **180**, 367–380 (2019).
32. Serrano-Bueno, G. *et al.* CONSTANS-FKBP12 interaction contributes to modulation of photoperiodic flowering in Arabidopsis. *Plant J. Cell Mol. Biol.* **101**, 1287–1302 (2020).
33. Garcia, A. V., Al-Yousif, M. & Hirt, H. Role of AGC kinases in plant growth and stress responses. *Cell. Mol. Life Sci. CMLS* **69**, 3259–3267 (2012).
34. Liu, Y. *et al.* The Maize ZmBES1/BZR1-9 Transcription Factor Accelerates Flowering in Transgenic Arabidopsis and Rice. *Plants* **12**, 2995 (2023).
35. Yin, Y. *et al.* Two interacting basic helix-loop-helix transcription factors control flowering time in rice. *Plant Physiol.* **192**, 205–221 (2023).
36. Hansen, C. C., Nelson, D. R., Møller, B. L. & Werck-Reichhart, D. Plant cytochrome P450 plasticity and evolution. *Mol. Plant* **14**, 1244–1265 (2021).
37. Phukan, U. J., Jeena, G. S., Tripathi, V. & Shukla, R. K. Regulation of Apetala2/Ethylene Response Factors in Plants. *Front. Plant Sci.* **8**, 150 (2017).
38. Carles, C. C., Choffnes-Inada, D., Reville, K., Lertpiriyapong, K. & Fletcher, J. C. ULTRAPETALA1 encodes a SAND domain putative transcriptional regulator that controls shoot and floral meristem activity in Arabidopsis. *Development* **132**, 897–911 (2005).
39. Zhao, M. *et al.* Natural variation of CsUGT71A60 determines growth and cold tolerance via regulating cytokinin glycosylation in Camellia sinensis. *Plant Biotechnol. J.* **23**, 2809–2823 (2025).
40. Abe, M. *et al.* FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. *Science* **309**, 1052–1056 (2005).
41. Wang, B. *et al.* UGT87A2, an Arabidopsis glycosyltransferase, regulates flowering time via FLOWERING LOCUS C. <https://nph.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-8137.2012.04107.x> (2012).

42. Li, Xu, Y. & Chong, K. The novel functions of kinesin motor proteins in plants. *Protoplasma* **249**, 95–100 (2012).
43. Amos, R. A. & Mohnen, D. Critical Review of Plant Cell Wall Matrix Polysaccharide Glycosyltransferase Activities Verified by Heterologous Protein Expression. *Front. Plant Sci.* **10**, (2019).
44. Covarrubias-Pazaran, G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLOS ONE* **11**, e0156744 (2016).
45. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
46. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.
47. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
48. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
49. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187-191 (2014).
50. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
51. Lopez-Cruz, M. *et al.* Multi-generation genomic prediction of maize yield using parametric and non-parametric sparse selection indices. *Heredity* **127**, 423–432 (2021).
52. Campos, G. de los, Sorensen, D. & Gianola, D. Genomic Heritability: What Is It? *PLOS Genet.* **11**, e1005048 (2015).
53. Pérez, P. & de los Campos, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **198**, 483–495 (2014).

54. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
55. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. & Cox, D. D. Hyperopt: a Python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **8**, 014008 (2015).
56. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at <https://doi.org/10.48550/arXiv.1705.07874> (2017).
57. Grabowski, P. P. *et al.* Genome-wide associations with flowering time in switchgrass using exome-capture sequencing data. *New Phytol.* **213**, 154–169 (2017).
58. Tornqvist, C.-E. *et al.* Transcriptional Analysis of Flowering Time in Switchgrass. *BioEnergy Res.* **10**, 700–713 (2017).