bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Prediction of plant complex traits via integration of multi-omics data

Peipei Wang[1,2,3#], Melissa D. Lehti-Shiu[3], Serena Lotreck[3,4], Kenia Segura Abá[1,5], Shin-Han Shiu[1,3,4,5#]

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

[2]Kunpeng Institute of Modern Agriculture at Foshan, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China

[3]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

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

[5]Genetics and Genome Sciences Program, Michigan State University, East Lansing, MI 48824, USA

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

The mechanistic bases of complex traits are consequences of activities at multiple molecular levels. However, connecting genotypes and these activities to complex traits remains challenging. We built prediction models using genomic, transcriptomic, and methylomic data for six Arabidopsis traits. Single data-based models performed similarly but identified different benchmark genes. In addition, distinct genes contributed to trait prediction in different genetic backgrounds. Models integrating multi-omics data performed best and revealed gene interactions, extending knowledge about regulatory networks. These results demonstrate the feasibility of revealing molecular mechanisms underlying complex traits through multi-omics data integration.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Translating genotypes to phenotype is challenging because the genetic mechanisms underlying trait variation are complex. Although genetic variation information is commonly used to predict phenotypes (i.e., genomic prediction)[1,2], researchers have had success in using other types of data. For example, transcriptomic data have been used to predict flowering time and yield[3] and pathogen resistance in plants[4]; methylomic data have been used to predict flowering time and plant height in a panel of epigenetic recombinant inbred lines of *Arabidopsis thaliana*[5,6]; metabolomic data have been used to predict biomass- and bioenergy-related traits in maize[7] and yield in rice[8]. Although multi-omics datasets that align with trait variation information are scarce in non-medical, multicellular model systems, the Arabidopsis 1001 Genome Project has generated phenotypic, genomic (G), transcriptomic (T), and methylomic (M) data for hundreds of accessions of the model plant *Arabidopsis thaliana*[9–11]. The availability of these datasets provides an opportunity to predict complex traits using machine learning approaches by integrating different data types. Through interpreting these machine learning models, gene features important for prediction of complex traits can be identified, to gain a deeper insight into the mechanistic basis of complex traits beyond the few significant quantitative trait loci (QTLs) that can be revealed through genome-wide association studies (GWAS).

Here we generated models for predicting Arabidopsis traits using G, T, and M data independently and combined. Although different single omics data-based models performed comparably in predicting six complex traits, they identified different subsets of benchmark genes when predicting flowering time. The combined model outperformed single omics data-based models, potentially because interactions between G, T, and M features were considered. In addition, we experimentally validated nine genes inferred to be important for predicting flowering time that were not previously reported to be involved in flowering. Our results highlight the utility of integrating multi-omics data in predicting complex traits and revealing underlying molecular mechanisms.

Results & Discussion

Prediction of complex traits using individual multi-omics data

To investigate the utility of multi-omics data for plant complex trait prediction, we compared the performances of models built using individual and combined data types. We first collected G

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(i.e., biallelic single nucleotide polymorphisms [SNPs]), T (i.e., RNA sequencing), and gbM (i.e., CG-, CHG-, CHH-type gene-body methylation) data and phenotypic values for six traits—flowering time, rosette leaf number (RLN), cauline leaf number (CLN), diameter of the rosette (DoR), rosette branch number (RBN), and stem length (SL)—for 383 Arabidopsis accessions from published studies[9–11] (**Table S1**). Across Arabidopsis accessions, G similarity (kinship) was more highly correlated with gbM similarity (mCor, Pearson Correlation Coefficient [PCC]=0.43) than with T similarity (eCor, PCC=0.18, **Fig. 1a**). In addition, G, T, or gbM data alone explained only a small amount of variation (PCC=-0.01–0.17, **Fig. 1a**), and there was no obvious relationship between trait and omics data similarities between accessions (**Fig. S1**), consistent with the findings in our previous study of yield, height and flowering time in maize[3]. This weak correlation is likely because only a subset of G/T/gbM variants (e.g., a few SNPs) are expected to contribute to the variation in a complex trait, whereas the linear correlation between the whole set (e.g., all SNPs in the G data) and complex trait variation is low.

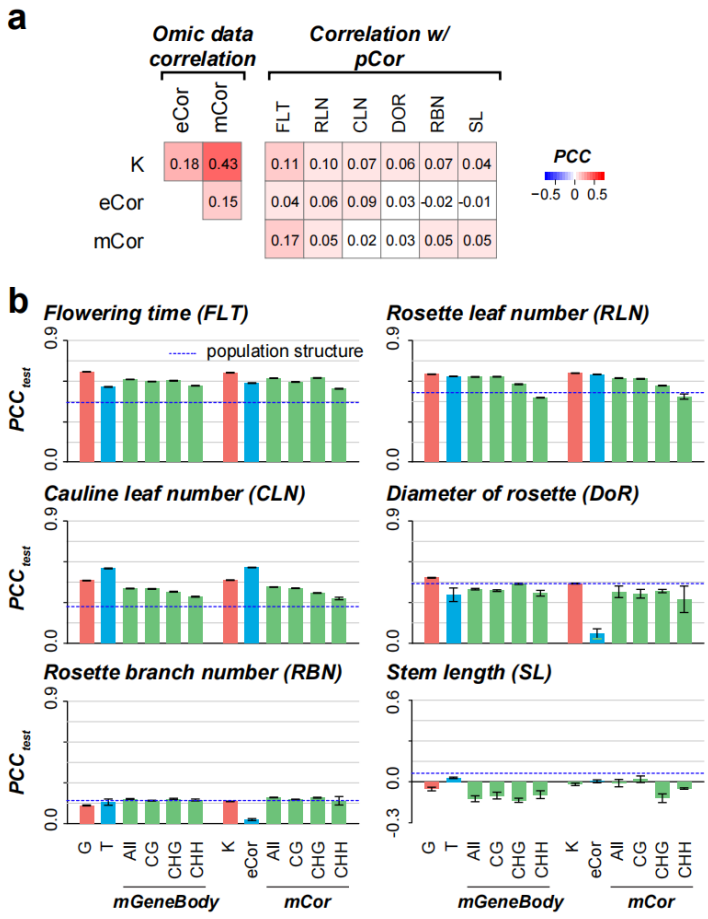


Fig. 1 | Relationships between omics data and traits and trait prediction model performance. a, Correlations between kinship (K), transcriptomic (expression) similarity (eCor), methylomic similarity (mCor), and phenotypic similarity (pCor) among Arabidopsis accessions. PCC: Pearson correlation coefficient. **b**, Performance of predicting models for six traits. Models were built using genomic (G), transcriptomic (T) and methylomic data (mGeneBody) or omics data similarities (K, eCor, mCor) among accessions with the ridge regression Best Linear Unbiased Prediction (rrBLUP) algorithm (for models built with Random Forest, see **Fig. S2a**). The PCC between true and predicted trait values in the test set (20%, held out before training the model) was used to evaluate model performance. The PCCs of models using population structure (first five principal components of genetic variation) are indicated by blue dashed lines. Error bars: standard deviation of 10 prediction runs.

While the overall correlations are low, the likelihood that predictive information is encoded within the overall G/T/gbM data led us to build machine learning models to take advantage of all

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

features from single omics data for trait prediction. We established single omics data-based trait prediction models using two algorithms, ridge regression Best Linear Unbiased Prediction (rrBLUP)[12] and Random Forest (RF)[13], and the model performance was assessed using a hold-out test dataset and measured as the PCC between true and predicted trait values (see **Methods**). As expected, the higher the correlation between the omics data and trait values (**Fig. 1a**), the higher the model performance was for most traits and omics data types (**Fig. 1b**, **Fig. S2a-d**).

For each trait, model performance was similar regardless of algorithm or whether omics data values or value similarities among accessions were used as predictive features (**Fig. 1b, Fig. S2a,e**). Most importantly, G-, T-, and gbM-based models had similar performances (**Fig. 1b, Fig. S2a**), consistent with the findings of our previous maize study using both G and T data for trait prediction[3]. Since CG, CHG, and CHH methylation have different regulatory mechanisms and functions in plants[14], we also established models using the different gbM types as separate features. Models using combined gbM data tended to have similar or better performances than models using individual gbM data types, and models using CHH methylation tended to have the worst performance (**Fig. 1b, Fig. S2a**). Because of the complexity of M data (e.g., heterogeneity in numbers of methylated cytosine sites within genic regions), we also explored six additional derived M features (single site-based M, hereafter referred to as ssM, see **Methods**) for predicting flowering time as an example. rrBLUP models using ssM features had higher prediction accuracies than those using gbM (**Fig. S3a**). This may be because the ssM features captured the distribution of methylation across each gene, which provided more detailed information about methylation patterns than the gbM data. Taken together, these results indicate that, similar to G and T data[3,15], M data are also useful for predicting plant traits, and that M data need to be represented in different ways to maximize their predictive power.

Distinct contributions of omics data to phenotype predictions

To identify the informative variants embedded in the G/T/gbM data, we investigated the importance of features for trait prediction by interpreting the prediction models. Three measures were used to evaluate feature importance: (1) coefficients of features in the rrBLUP models (**Fig. S4a-c**); (2) gini importances in the RF models (**Fig. 2a-c**); and (3) the average absolute SHapley Additive exPlanations (SHAP) values[16] obtained from RF models (**Fig. S4d-f**). Here we focused on features important for predicting flowering time because there is abundant knowledge about the genetic control of this trait, which is crucial for interpreting the important features. To allow comparison with T and gbM features, which were gene-based, G variants were mapped to genic regions (see

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods). Genes corresponding to or harboring important G/T/gbM features (defined as those with >95th percentile importance values) were considered important for flowering time prediction and are hereafter referred to as important genes. We found weak or no correlation between importance scores from models built using different types of omics data, regardless of the feature importance measure examined (ρ: -0.07–0.11), and there was little overlap of important genes between models (**Fig. 2, Fig. S4a-f**).

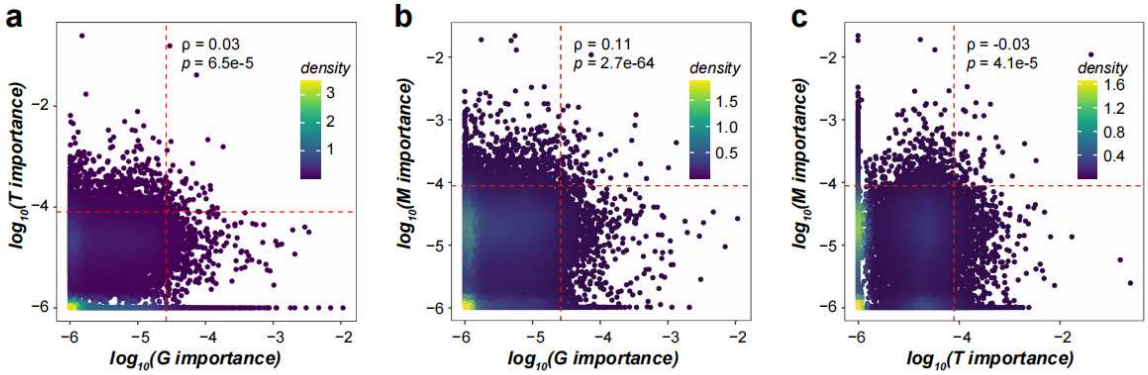


Fig. 2 | Correlation of feature importance between different types of omics data. a-c, Density scatterplot of the importance scores between genomic (G), transcriptomic (T), and methylomic (M) features of genes when RandomForest gini importance scores were used as measures of importance. **a**, G vs T, **b**, G vs M, **c**, T vs M. All feature importance < 10^{-6} was assigned to be 10^{-6}. Red dashed line: 95th percentile of feature importance. Density plot color: density of genes. ρ: Spearman's rank correlation coefficient.

These results suggest that the similar trait prediction accuracy of models built with different omics data is not due to shared features, consistent with the findings of the maize study[3]. However, the correlation between the importance of G and gbM variants (**Fig. 2b, Fig. S4b,e**) was higher than that for other comparisons, consistent with the relatively higher correlation between G and gbM similarities across accessions (**Fig. 1a**) and suggesting potential confounding effects of G on gbM data. To disentangle gbM from G data, we built trait prediction models with the mCor residuals to exclude the G similarity (kinship) effects (**Methods**). Consistent with a confounding effect of G on gbM, these new models had significantly lower performance compared with models based on the mCor matrix for all traits (differences in PCC scores used as a measure of prediction accuracy: 0.06–0.18, p<0.001) except RLN and CLN (differences: 0.03 and 0.01, p=0.05 and 0.09, respectively, **Fig. S4g**). However, mCor-based models with confounding effects of kinship removed still performed better at predicting flowering time and RLN than those using the first five principal components to approximate population structure (**Fig. S4g**). The performance of the population structure-based model is used as the baseline for trait prediction. Thus, the components of gbM independent from G are important for trait prediction.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Benchmark flowering time genes identified as important features in flowering time prediction models

In the previous section we showed that models built using different omics data identified different important genes (**Fig. 2, Fig. S4a-f**). We next asked how many genes with known functions in flowering time were identified as important genes. To do this, we obtained 378 benchmark flowering time genes (**Table S2**) from the database FLOR-ID[17] and found that 148 of these genes were identified as important according to at least one of the three importance measures for at least one of the three individual omics datasets (**Fig. 3**, **Table S3**). Only two genes, *FLOWERING LOCUS C* (*FLC*) and its paralog *MADS AFFECTING FLOWERING 2* (*MAF2*), were identified as important by all three independent omics datasets (blue, **Fig. 3**). This is consistent with the roles of *FLC*[10,18,19] and *MAF2*[20,21] in flowering time regulation being established through studies of genetic variation, transcript levels, and methylation levels. Another 23 genes (green, **Fig. 3**) were identified by two independent omics datasets. For instance, *FLOWERING CONTROL LOCUS A* (*FCA*), which increases H3K4 dimethylation in the central region of *FLC* and regulates its expression[22], was considered important in the G and gbM models. The remaining 123 genes were dataset specific (**Fig. 3**, **Table S3**), such as *SUPPRESSOR OF OVEREXPRESSION OF CO 1* (*SOC1*), which was only identified as important in T models. This is consistent with our observations that there is little overlap between important genes identified by G, T, and gbM models (**Fig. 2a-c, Fig. S4a-f**), and that gbM data, although it is confounded with G information (**Fig. S4g**), can make unique contributions.

We next asked whether significantly more benchmark flowering time genes than random chance were identified by our models. When the 95th percentile was used as the threshold, only G and T models with RF gini importance scores identified significantly higher proportions of benchmark genes (8.84% and 8.06%, respectively) than expected by random chance (p=0.002 and 0.014, respectively, Fisher's exact test. **Table S4**). To understand why no more benchmark genes were identified than expected by random chance for most models and importance measures, we explored the following potential factors (for the rationale and analysis process, see **Methods**): cut-off threshold (95th or 99th percentile), confidence of benchmark gene functions in FLOR-ID, difference in gene contributions to flowering in short days and/or long days, and the number of accessions used for training models. All these factors had little impact on the identification of benchmark genes (**Table S4-S6**), except for SHAP values when the 99th percentile was used, where significantly more benchmark genes than expected by random chance were identified for the T model (**Table S5**) .

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

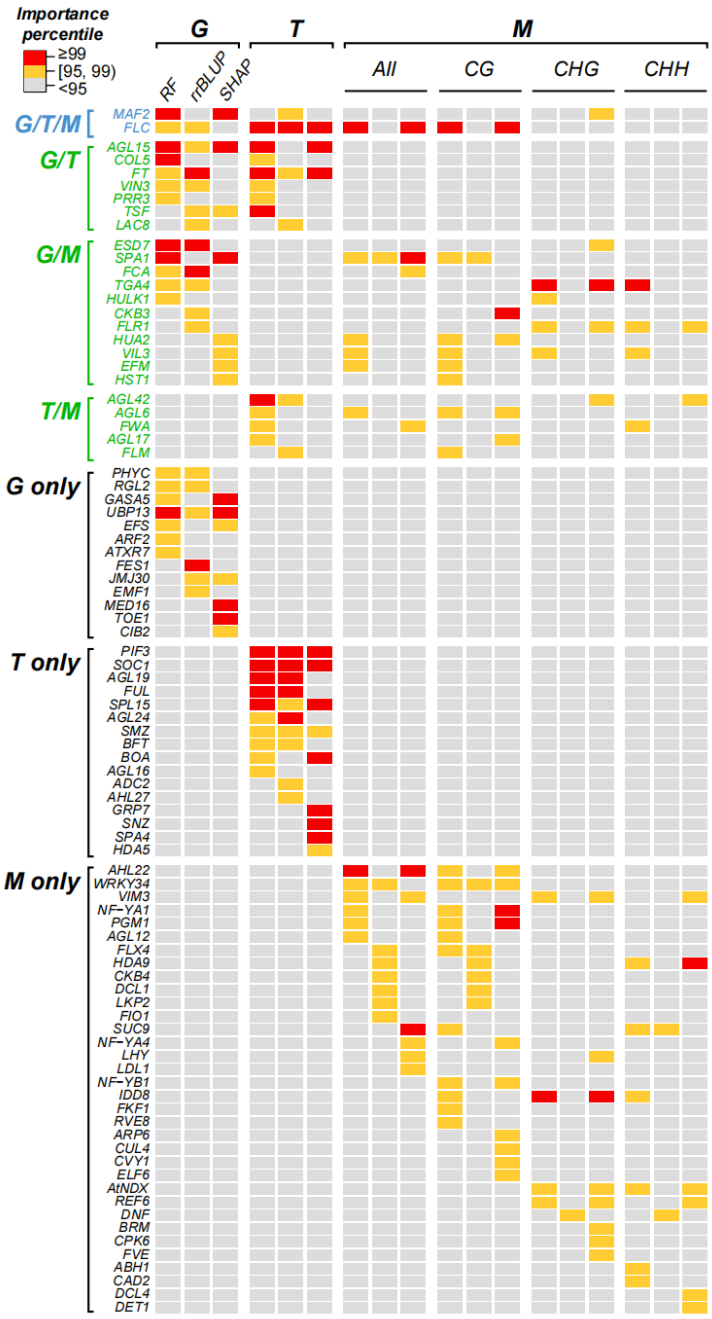


Fig. 3 | Example benchmark flowering time genes identified using different importance measures and datasets. The heatmap shows how often benchmark flowering time genes were identified as important using different datasets.For each omics dataset, three feature importance measures are used: (1) RF - Random Forest gini values; (2) rrBLUP - absolute value of rrBLUP coefficient; (3) SHAP - SHapley Additive exPlanations values. Color in the heatmap indicates importance value percentile: ≥99th percentile, red; ≥95th and <99th percentile, orange; <95th percentile, gray. Font color indicates the number of omics datasets that identified a benchmark flowering time gene as important: all datasets, blue; two datasets, green; a single dataset, black. G, T, and M: genomic, transcriptomic, and methylomic data, respectively.

One thing worth noting is that gbM-based models identified no more benchmark genes as important than expected by random chance, regardless of the threshold used. This may be because the gbM does not adequately represent the methylation profiles of a given gene. To assess this, we also interpreted the ssM-based models where the methylation profile across a gene region was represented (see **Methods**). An additional 122 benchmark flowering time genes were recovered as important by at least one ssM feature type/feature importance measure combination (**Table S3**), and more benchmark genes were identified as important for a single ssM model (a maximum of 42

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

genes, **Table S7**) than gbM models (a maximum of 22 genes, **Table S4**). One example gene is *CLEAVAGE STIMULATION FACTOR 77* (*CSTF77*), which is involved in the 3' processing of antisense *FLC* mRNA[23]. Accessions with CG-type methylation at two different sites in *CSTF77* had significantly longer flowering times than those without (**Fig. S5a,b**), but there were no significant differences in flowering time between accessions with different SNP alleles (**Fig. S5c,d**). In addition, there was no correlation between the expression levels or gbM levels of *CSTF77* and flowering time (**Fig. S5e-h**). These results explain why *CSTF77* was uniquely identified by ssM-based models (**Table S3**).

Another potential reason for the low degree of enrichment of identified benchmark genes for most models is the fact that the plants scored for flowering time phenotypes were grown at 10 °C and those used for the G, T, and M data were grown at 22 °C. We also built models using the same G/T/M data to predict flowering time phenotypes recorded at 16 °C[10], and found that the temperature at which flowering time was scored affected gene importance for flowering time prediction (**Table S8, Fig. S6a**). For example, the expression of *FRIGIDA* (*FRI*), which, when functional, confers a vernalization requirement[24], had a SHAP value of zero for flowering time prediction at 10 °C but a SHAP value of 0.075 (ranked 14th) at 16 °C. This is consistent with the larger differences in flowering time between accessions with functional and non-functional *FRI* copies at 16 °C than at 10 °C (a temperature at which the vernalization requirement is met for accessions with functional *FRI*); the larger SHAP values for *FRI* in the 16 °C model indicate that it makes a higher contribution to flowering time prediction when the plants are unable to vernalize (**Fig. S6b**). These results indicate that different omics datasets, ways to represent data, importance measures, and environmental factors must be considered to better capture the genetic basis of Arabidopsis flowering time and likely other complex traits.

Identification of novel genes involved in regulating flowering time

To determine whether all the features relevant to benchmark flowering time genes are sufficient to predict flowering time, we built RF models using G, T and gbM features for 378 benchmark genes separately or combined (hereafter referred to as benchmark gene-based models). Compared with the corresponding full models built using the features for all genes, the benchmark gene-based models had significantly lower performance (**Fig. 4a**), suggesting that genes in addition to the 378 benchmark genes are involved in regulating flowering time. To test this, we selected the 378 most important genes that were not benchmark genes (hereafter referred to as "novel" genes) from the full model, which was built using G/T/gbM features combined for all genes (**Table S9**). We found

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

that the novel gene-based gbM model (built using the gbM features of these novel genes) performed significantly better than the benchmark gene-based gbM model, but not the corresponding G and T models (**Fig. 4a**). This may be attributed to the fact that most benchmark flowering time genes were identified using genetic approaches (e.g., via forward genetic screens or GWAS) and/or transcriptomic data (e.g., via gene differential expression analysis), rather than methylomic data.

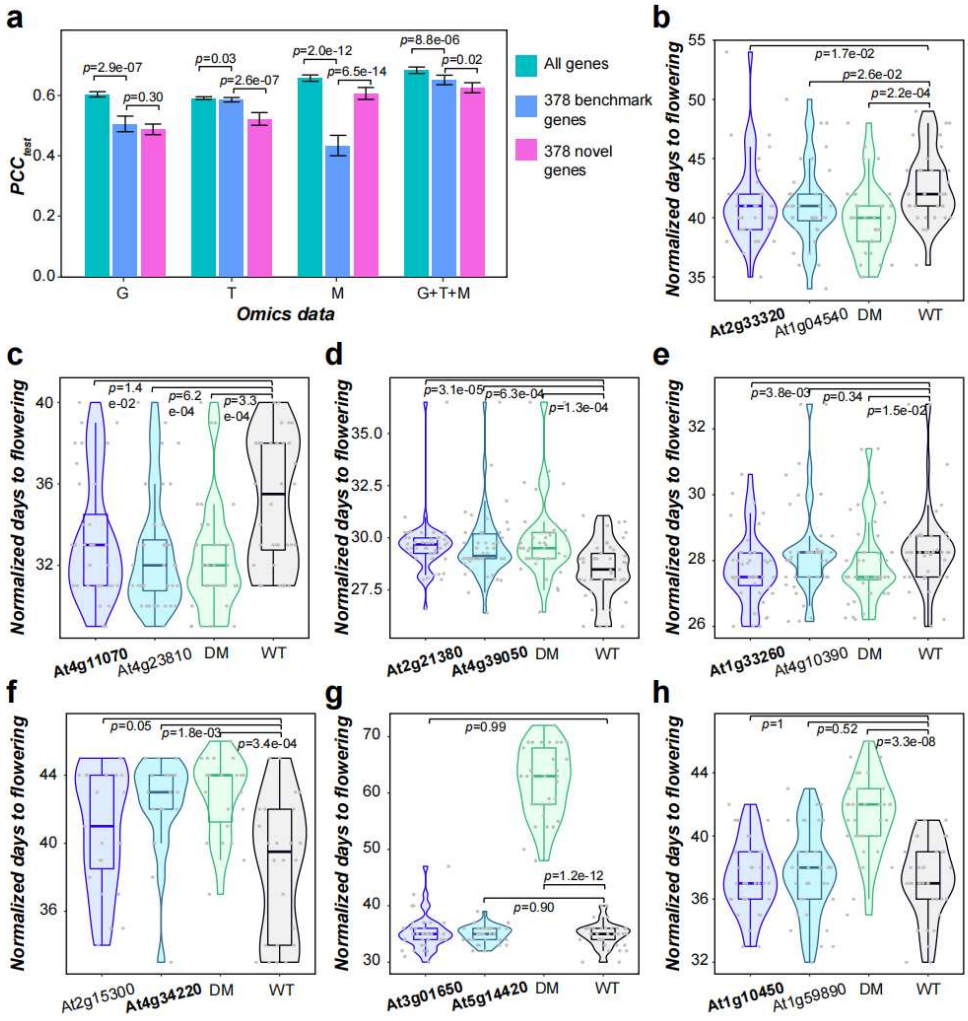


Fig. 4 | Identification of novel genes involved in flowering time regulation. a, Accuracy of RF flowering time prediction using all features (green), only features related to benchmark flowering time genes (blue), or only features related to the top 378 novel genes (magenta). G, T and M: genomic, transcriptomic and methylomic data, respectively. *p*-values are from Student's t-test. **b-h**, Statistical analysis of flowering time in single-gene mutants of important genes (bold font) and their paralogs (regular font), double mutant (DM) with loss of function of both paralogs, and wild-type (WT) plants. *p*-values are from Wilcoxon Rank Sum test. Flowering time was normalized across flats within blocks (see **Methods**). Horizontal line in the box: median value; box range: interquartile range (IQR), 25th (Q1) to 75th percentile (Q3); whisker below box: Q1–1.5 IQR to Q1; whisker above box: Q3 to Q3+1.5 IQR; violin plot: distribution of datapoint values; dot: datapoint from an individual plant.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The findings that the full models and novel gene-based gbM model outperformed the benchmark gene-based models are consistent with the hypothesis that these novel genes are likely involved in regulating flowering. Through a literature search, we found that four of the 378 novel genes, including *DELAY OF GERMINATION 1* (*DOG1*)[10,25], a SNP of which had a importance percentile > 0.99 (**Table S10**), have been shown to play roles in flowering time regulation even though they were not included as benchmark genes in the FLOR-ID database (**Table S10**). To validate the functions of important genes in flowering time, we took advantage of an existing dataset in our lab to compare flowering time in mutants of 21 non-benchmark genes that were identified as important and the wide type (WT) (**Methods**, **Table S11**,**12**). Six of these 21 genes affected flowering time when mutated (**Table S11**; **Fig. 4b-h**). For example, a loss-of-function mutant of *AT4G11070* (*WRKY41*), which controls seed dormancy[26], flowered significantly earlier than WT (**Fig. 4c**). Consistent with this, *WRKY41*'s homolog *Dlf1* regulates flowering in rice[27]. Another three genes had loss-of-function effects on flowering when mutated along with their paralogs (**Fig. 4g,h**, **Table S11**). The remaining 12 genes did not have a significant effect on flowering time when mutated, alone or with their paralogs. One potential explanation is that our validation experiments were conducted in the Col-0 genetic background, but the important genes were identified in models built across multiple accessions.

To check whether our models perform well in identifying genes that function in flowering time, we also measured the flowering time of the loss-of-function mutants of 37 "non-important" (importance rank ≤95th percentile) genes and their paralogs. We found that 43.2% (16 genes) had significantly altered flowering time when mutated (**Table S11,12**). This percentage is only slightly lower than that of experimentally validated important genes (52.2%). As discussed above, we explained why no significantly more benchmark genes can be identified due to e.g., dataset and environmental factors. In addition, there are only 383 accessions in our models that are fitted with far more features (e.g., >20,000 T features). Therefore, we do not have enough power to detect variants with significant but lower degrees of contribution to flowering time. Our false negative rate when using importance rank as a criterion is expected to be high.

Accession-dependent contributions of genes to flowering time prediction

Thus far, we have evaluated the contribution of G, T, and gbM features associated with genes to the prediction of flowering time across accessions by dissecting the models through global interpretation[28]. However, some genes may be important contributors to flowering time only in specific accessions. To assess this, we determined the contributions of important features to

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

flowering time in each accession (local interpretation) by examining the SHAP value for each feature in each accession (individual SHAP values, as opposed to the averaged value discussed earlier, see **Methods**). Here, a positive SHAP value for a feature in an accession means that the value of the feature in that accession contributed to a higher predicted trait value, i.e., longer flowering time. A negative SHAP indicates the opposite: the feature value contributed to reduced flowering time in an accession. The absolute SHAP value describes the degree of feature contribution to trait prediction. Here we first present the SHAP values of the top 20 important genes from the T model in detail as an example because more benchmark genes were among the top 20 genes in this model (**Fig. S7a**) than in the G (**Fig. S7b**) and gbM models (**Fig. S7c**). Organizing the accessions into clusters based on SHAP values of the top 20 T features allowed us to examine the way that different features contribute to flowering time prediction. The accessions we examined formed eight clusters(**Fig. 5a**), and flowering time varied greatly across these clusters (**Fig. 5b**).

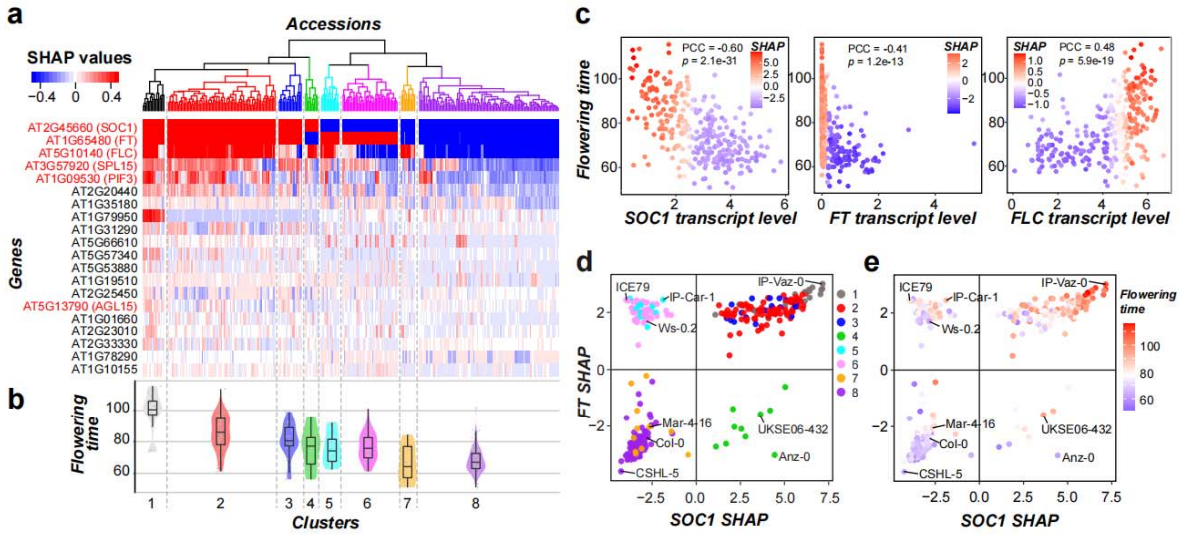


Fig. 5 | Accession-dependent effects of transcriptome features on flowering time. a, Heatmap shows the SHAP values of top 20 genes (y-axis) for each accession (x-axis). Benchmark gene names are in red on the left of the heatmap. **b**, Violin plot shows distribution of flowering time of accessions within each cluster. Horizontal line in the boxplot: median value; box range: interquartile range; dot: individual accession. **c**, Correlation between *SOC1* (left), *FT* (middle) and *FLC* (right) expression levels with flowering time. Color scale: SHAP values for *SOC1*, *FT* and *FLC* when using their expression to predict flowering time. PCC: Pearson correlation coefficient, showing correlation between gene expression (x-axis, log[TPM + 1]) and flowering time (y-axis). **d-e**, Relationships between *SOC1* and *FT* SHAP values among accessions with example accessions labeled. Colors in **d**: memberships of accessions in clusters specified in **a**; color scale in **e**: flowering time.

In cluster 1 and 2 accessions (**Fig. 5a,b**), *SOC1*, *FT*, and *FLC* expressions all have positive SHAPs, indicating they contribute to longer flowering time. Their positive contribution is mainly due to lower *SOC1* and *FT* expression and higher *FLC* expression compared with other accessions (**Fig. 5c**). In cluster 8 accessions (**Fig. 5a,b**), all three genes have negative SHAPs, and the shorter

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

flowering time is due to higher *SOC1* and *FT* expression but lower *FLC* expression. This is consistent with earlier findings that *SOC1* and *FT* promote flowering[29], *FLC* represses flowering[30], and the expression levels of *SOC1* and *FT* are negatively correlated with *FLC* expression[31]. This coupling between *SOC1* and *FT* and *FLC* in terms of expression (**Fig. 5c, Fig. S8a,b,f,g**), SHAP values (**Fig. 5d, Fig. S8c**), and flowering time contribution (**Fig. 5e, Fig. S8e**) is the predominant flowering time regulatory mechanism for most accessions. However, the action of these components can be decoupled. For example, in cluster 4, 5, and 6 accessions (green, cyan, and magenta; **Fig. 5d**), the SHAP values of *SOC1* and *FT* have opposite signs, which is associated with moderate flowering time values (**Fig. 5b,e**). Another example is cluster 7 (orange, **Fig. 5d**) where, despite the positive contribution of *FLC* (**Fig. S8c**), both *SOC1* and *FT* contribute negatively to flowering time (**Fig. 5d**). This coupling and uncoupling between *SOC1*, *FT* and *FLC* expression across accessions indicates that regulation of flowering time may be more complex than what has been reported. Accession-dependent contributions of other important T features are also observed, although to a much lower extent. An example is *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 15* (*SPL15*); clusters 2 and 4–8 can be divided into sub-clusters depending on whether *SPL15* expression contributes positively to flowering time in an accession (**Fig. 5a**).

We also clustered accessions using SHAP values of the top 20 important features from G and gbM models, and obtained eight and nine clusters, respectively (**Fig. S9a,c**), which resembled the distribution of T model-based clusters in that the clusters could be characterized by a few top features. For example, the second (*AT4G38550*) and the third (*AT1G35610*) most important genes from the gbM model can be coupled or uncoupled with *FLC* in a similar fashion as *SOC1* and *FT* (**Fig. S9c**). The different omics data types yielded clusters with different accessions (**Fig. S10**). These findings further demonstrate that different omics data reveal different contributors to flowering time variation among accessions, and that the different effects of genes on flowering time among accessions can be disentangled through model interpretation. In addition, knowledge about flowering regulation in some accessions may not be generalizable to others, as demonstrated by the identification of different benchmark genes when different sets of accessions were used (**Table S4,6**). This may be explained by the differences in genetic backgrounds among accessions[24] and the high complexity of genetic interaction networks regulating flowering[32]. The accession-dependent effects of genes on flowering may also partially explain the low degree of enrichment of benchmark genes among important genes in our original G, T and gbM models because these benchmark genes were predominantly discovered in the Col-0 accession.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genetic interactions revealed through integration of multi-omics data

In above sections we showed that different types of omics data revealed overlapping but mostly distinct genes impacting flowering time. Next, we asked whether combining different types of omics data improves the prediction accuracy. We found that combining G and T or all three datasets improved model performance for RF models, but not for rrBLUP models (**Fig. 6a**). Because the RF algorithm considers non-linear feature combinations while rrBLUP does not, the better RF model performance suggests that the inclusion of interactions between features from different omics data may have improved model performance. To evaluate this possibility, we established an additional RF model, which only included G+T+gbM features relevant to the 378 benchmark flowering time genes (see **Methods**) to facilitate model interpretation. We used the SHAP approach[33] to identify feature interactions (see **Methods**), where the contribution of feature X to trait prediction is influenced by values of feature Y. The SHAP feature interaction can help us identify potential genetic interactions between genes or variants, such as epistasis[34].

We identified 65,829 feature interactions, including all six possible combinations between omics data types: G-G, T-T, gbM-gbM, G-T, G-gbM, and T-gbM (**Table S13**). T-gbM and T-T interactions were the most prevalent and had the highest interaction values (**Fig. 6b**). The four most important genes in the T model—*SOC1*, *FT*, and *FLC*, and *SPL15* (**Fig. 5a, Fig. S7**)—had the highest numbers of feature interactions (1007, 989, 813, and 787, respectively), and had interactions between each other that ranked above 20th according to the interaction values (**Fig. 6c**, **Fig. S11a-f**, **Table S13**). Interactions between *SPL15* with the other three genes have been reported in the literature: *SPL15* is repressed by *FLC*[35] and acts in parallel with *SOC1* to activate *FRUITFULL* (*FUL*) expression[36]. Examples for other feature interactions were shown in **Fig. 6d-f** and **Fig. S11**, including these between features of gene pairs that have known biological interactions. For example, *AGAMOUS-LIKE 15* (*AGL15*) was reported to bind to the *SOC1* promoter[37](T-T interaction between *AGL15* and *SOC1* ranked 11th, **Fig. 6d**); *NUCLEAR FACTOR Y SUBUNIT A1* (*NF-YA1*) binds to the *SOC1* promoter where it modulates H3K27me3 levels[38] (gbM-T interaction between *NF-YA1* and *SOC1* ranked 19th, **Fig. 6e**); *MEDIATOR 12* (*MED12*) up-regulates the expression of *FT*[39] (G-T interaction between *MED12* and *FT* ranked 8th, **Fig. S11n**).

Feature interactions for gene pairs with no reported biological interactions were also identified (referred to as novel feature interactions hereafter), and they might indicate potential biological interaction between genes. For example, T-G interaction between *SOC1* and *MADS AFFECTING FLOWERING 4* (*MAF4*) ranked 13th (**Fig. 6f**); although no direct biological interaction

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

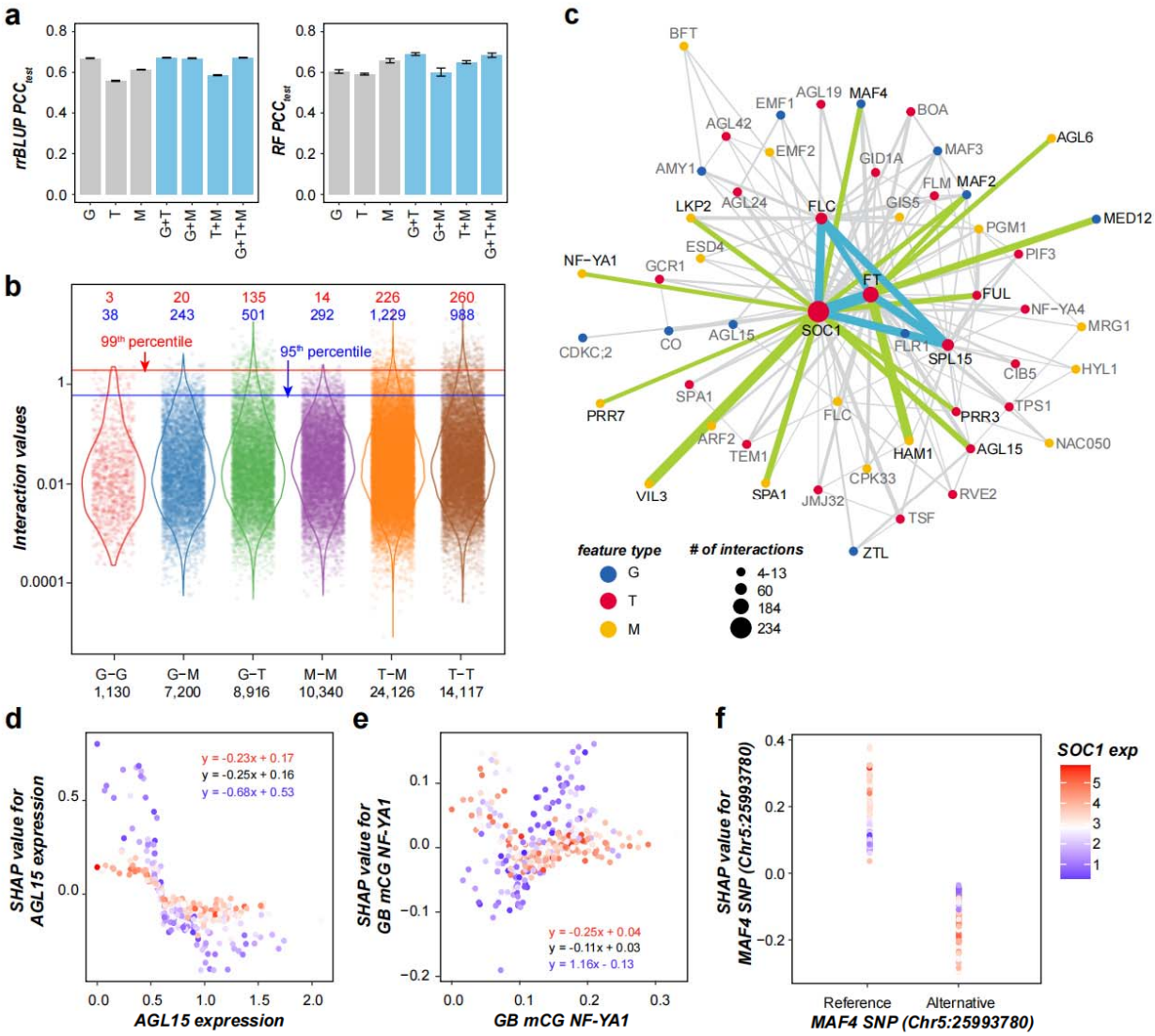


Fig. 6 | Prediction models integrating all omics data types and putative interactions between flowering time genes. a, Prediction accuracy of flowering time for models built using single and combined omics data. Left panel: rrBLUP models, right panel: RF models; G, T and M: genomic, transcriptomic and methylomic data, respectively; error bars: standard deviation of 10 runs. **b**, Distributions of SHAP interaction values for different interaction types. X-axis label: interaction type and number of identified interactions. Red and blue dashed lines: 99th and 95th percentiles of interaction values, respectively; numbers in red and blue fonts: numbers of feature interactions above the 99th and 95th percentiles, respectively for each type of interaction. **c**, Network of features with ≥ 4 interaction values above the 99th percentile. Blue, red and orange points: G, T and M features, respectively; node size: number of interactions; edge thickness: $\log_2$(interaction value/10 + 1); blue edges: interactions between expression of *SOC1*, *FT*, *FLC*, and *SPL15*; green edges: interactions among the top 20 interactions other than those connected with blue edges; black font: features involved in the top 20 interactions; gray font: other features. **d-f**, Interaction between *SOC1* expression and *AGL15* expression (**d**), *NF-YA1* CG-type gene-body methylation (**e**) and a *MAF4* SNP (**f**). Equations in **d,e** show regression of feature values (x-axis) and the SHAP values (y-axis) in accessions with *SOC1* expression levels (log[TPM + 1]) > 4 (red font), between 2 and 4 (black), and < 2 (blue).

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

between *MAF4* and *SOC1* has been reported, two close homologs of *MAF4*, *FLC* and *MAF2*, were shown to repress *SOC1* expression[40]. Another example is G-gbM Interaction between *EMBRYONIC FLOWER 1* (*EMF1*) and *GIBBERELLIN 20-OXIDASE 3* (*GA20OX3*) ranked 266th (**Fig. S11r**); biological interaction between *EMF1* and *GA20OX3* has not been reported before. Taken together, these findings show that potential interactions at different molecular levels can be identified on a large scale by interpreting computational models integrating multiple omics data.

Conclusion

We investigated the utility of G, T, and M data in predicting complex traits in Arabidopsis. Models built using these data types separately had similar performances, but identified different benchmark flowering time genes and other genes not previously reported to be involved in flowering. Even models built using different forms of a single type of omics data—M data—identified different sets of flowering genes. These results highlight the necessity of exploring different types of omics data when predicting complex traits. Even though there was only one time point for transcriptome data (rosette leaves right before bolting)[9], key regulators of flowering time, such as *FLC*, *FT*, *SOC1*, and *SPL15*, were still identified in T-based rrBLUP and RF models. Considering that rrBLUP is an algorithm based on a linear model, this finding indicates that simple, linear combinations of the steady-state expression levels of these genes that explain a significant part of the flowering time variation among accessions can be identified.

We identified important genes including the most well known flowering time genes; however, the high false positive and false negative rates based on the benchmarks were high. The model performance and candidate gene identification can be further improved by incorporating significantly more accessions, using additional approaches to select, combine, and represent features, and incorporating data from more than one environment. In addition, the Arabidopsis T and M data used in this study were obtained from mixed rosette leaves harvested just before bolting, which poses potential problems for complex trait prediction: 1) gene expression and methylation can be cell-type specific, thus noise in T and M data is inevitable due to cell heterogeneity; 2) the complex traits to be predicted may be specific to certain tissues, organs, or development stages; thus, T and M variation from leaves may not reflect trait variation in other contexts. The development of single-cell and spatial omics techniques, which allow the cellular landscape of epigenomes within a tissue to be measured, is expected to improve the prediction accuracy of complex traits. In addition, omics data in addition to G, T, and M data can be used to

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

predict complex traits; these data include chromatin architecture, chromatin accessibility, and histone modification, changes of which have been shown to regulate flowering in Arabidopsis[41–44].

That the effects of genes or variants on flowering differ among accessions is known in specific cases; for example, DNA demethylation has different effects on flowering in C24 and Landsberg *erecta*[45], and some accessions carry non-functional alleles of the *FRI* gene[24]. Here, we show that accession-specific effects of different flowering time genes can be revealed by interpreting machine learning models. The outcome also revealed the accession-dependent effects that were not documented previously. Interpretation of a non-linear model (RF) also allowed the identification of known and novel interactions among G, T, and gbM features of benchmark genes. The novel interactions represent new hypotheses that require further experimental verification and are expected to provide insights into how these novel components and interactions contribute to the genetic basis of flowering time and potentially other complex traits.

Methods

Data preprocessing

Six traits for each Arabidopsis accession (**Table S1**) were obtained from two publications: (1) flowering time at 10 °C or 16 °C, which was scored as days until the first flower was open, was from[10]; (2) cauline leaf number (CLN), (3) rosette leaf number (RLN), (4) rosette branch number (RBN), (5) diameter of rosette (end point, after flowering, DoR), and (6) length of main flowering stem (stem length, SL) were from[11]. For each trait, the pairwise Euclidean distances of trait values between accessions were calculated using the R package “rdist”. The Euclidean distances were first normalized between 0 and 1, and then the similarity of trait values among accessions was calculated as 1 - normalized Euclidean distance.

The genomic matrix (G) was downloaded from the 1001 Genomes database[10] (http://1001genomes.org/data/GMI-MPI/releases/v3.1/SNP_matrix_imputed_hdf5/). This matrix only contains biallelic SNPs. SNPs with minor allele frequency < 0.05 were removed. For each SNP, the major allele was encoded as 1, and the minor allele was encoded as -1. The kinship matrix and the first five principal components from SNPs were generated using the centered Identity By State method[46] implemented in TASSEL v5[47].

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For the transcriptomic (T) data[9], the normalized read counts were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE80744), and were used to calculate the transcripts per million (TPM) using the function "calculateTPM" from the R package "scater"[48]. The TPM values were then transformed to ln(TPM+1). The transformed TPM values were used to calculate the expression correlations (eCor) among accessions, and the resultant Pearson correlation coefficient (PCC) values were normalized between 0 and 1 to make them comparable with pCor, which ranges from 0 to 1.

For the methylomic (M) data[9], the gene-body methylation values (the number of reads with methylated cytosines in a gene body divided by the total number of reads with both cytosines and thymines) were downloaded from http://signal-genet.salk.edu/1001.php. For accessions with multiple replicates, the median values were calculated across replicates. The missing values were imputed using the k-Nearest Neighbors approach ("KNNImputer" function in sklearn.impute)[49]. The imputation was conducted five times with different k values (3, 4, 5, 6, and 7), and the means of imputed values were used. The imputed gene-body methylation values were used to calculate the gene-body methylation correlation (mCor) among accessions, and the resultant PCC values were normalized between 0 and 1.

Additional formats of methylomic data

The frequency of cytosine sites can vary within a genic region, and the methylation level at each cytosine site (i.e., proportion of mapped reads with methylated cytosine) can also vary because of cell heterogeneity; therefore, the gene-body methylation obscures information such as heterogeneity in methylation levels over the length of a gene. To overcome these shortcomings, we used six more types of M data (**Fig. S3**) to represent single site-based methylation (ssM) for predicting flowering time: (1) methylation status of single cytosine sites (i.e., 1, indicating methylated, or 0, indicating not methylated, hereafter referred to as presence/absence [P/A] of methylation) across the whole genome; (2) methylation proportion (i.e., proportion [Prop] of reads that were methylated at that site) for single cytosine sites across the whole genome; (3,4) methylation profiles along each gene, i.e., the methylation status (3, mean_P/A) or proportion (4, med_Prop) across upstream, gene-body, and downstream regions divided into 30 bins; (5,6) clusters of genes with similar methylation profiles when methylation status (5, P/A_clu) or proportion (6, Prop_clu) were considered.

The methylation base calls for each reference site[9] were downloaded from NCBI (accession ID GSE43857). Missing values in the single cytosine site methylation matrix were treated as

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

follows: (1) if the reference site was cytosine and there was a SNP for this site in accession X, or if the reference site was not a cytosine and this site was not a SNP for X, the methylation call for this site in X was encoded as 0; (2) if the site's base in accession X was unknown, or the site in X was a cytosine, but there were no reads mapped to this site, then the values were marked as missing to be imputed in the same way as for the gene-body methylation matrices, except that for the P/A of methylation for a single site, the final imputed value was rounded to either 0 or 1. After imputation, sites with the same value (either 0 or 1) in > 95% accessions (similar to 5% minor allele frequency) were removed. To balance the trade-off between the number of methylation sites included in the analysis and the number of missing data points, we first ordered the sites according to the proportion of missing values across accessions. Three thresholds were explored: 90^{th}, 75^{th} and 50^{th} percentile (the corresponding datasets are referred to hereafter as 90per, 75per and 50per, respectively), sites above which had missing values in ≤ 2, 8, and 38 accessions, respectively (**Fig. S12a**). We found that the dataset 50per tended to have the highest performance in rrBLUP models while dataset 90per led to the highest performance in RF models (**Fig. S12b,c**). To include more single-site methylation information, matrices using the 50^{th} percentile as a threshold were used for subsequent analysis.

To bin the methylation levels of single sites for gene-based regions, the gene regions were split into 30 bins as shown in **Fig. S12d**. For each bin, a summary statistic representing the methylation data was calculated: for the methylation P/A, this statistic is the mean number of methylated cytosines in the bin; for methylation proportion, this statistic is the median of the methylation proportion values (a number between 0 and 1) in a bin. To summarize the methylation profiles of genes across accessions, the binned methylation data were formatted into vectors, with one vector of length 30 (the number of bins) for each gene in each accession. These vectors were clustered using K-means clustering with 30 clusters. For each gene in each accession, 30 new features (i.e., whether the vector of the gene [methylation profile] belongs to each of the 30 clusters, 1 for the cluster the gene belongs to, 0 for all the other clusters) were produced. Finally, for each feature matrix, columns containing only zeros were removed. These processes were performed on each methylation type separately (CG, CHG, CHH), and matrices of three methylation types were combined together, resulting in 16 final feature matrices (16 "binned" columns in **Table S3**).

Predictive modeling

The R package "rrBLUP"[12] and the sklearn class "RandomForestRegressor"[13,49] were used to build the predictive models. Twenty percent of the accessions were randomly held out as the test set,

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

which was used to evaluate the performance of final models and was never used in the model training. The remaining 80% of accessions were used to train the models. For both model architectures, a five-fold cross-validation scheme was conducted in the model training. First, accessions were randomly split into five folds. Next, accessions in four folds (referred to as the training set) were used to build the model, and accessions in the fifth fold (validation set) were used to evaluate the model performance. Finally, this training-validation step was conducted five times to make sure each fold was used as a validation set once. This five-fold cross-validation scheme was repeated 10 times. The PCC was calculated between true trait values and the predicted values, and the average PCC among the 10 runs was used to measure the model performance on the cross-validation set.

For the RandomForestRegressor, GridSearch was used for hyperparameter tuning, where spaces of the parameters -max_depth and -max_features were [3, 5, 10], and [0.1, 0.25, 0.5, 0.75, “sqrt”, “log2”, “None”], respectively. The parameter -n_estimators was set as 100. The best parameter combination was selected according to the model performances in cross-validation. A final model was built using all the accessions (including accessions in both the training and validation sets) with the best parameter combination, and was applied on the test set to evaluate the model performance. For rrBLUP, the coefficients of features were estimated within each fold of cross-validation and were used to predict trait values for accessions in the validation and test set. After five folds of cross-validation, the prediction accuracies on the test set were averaged to evaluate the model performance.

To establish the baseline for the genomic prediction, we built predictive models based simply on population structure (defined as first five principal components from genetic markers, blue dotted line in **Fig. 1b**). Models based on individual omics data outperformed those based on population structure for flowering time, RLN, and CLN, but not for DoR, RBN, and SL. In addition, model performances for models based on individual omics data and population structure on the cross-validation (CV) and test sets for DoR differed dramatically in both rrBLUP and RF models (**Fig. 1b**, **Fig. S2**), which is indicative of high heterogeneity in DoR among accessions; the accessions in the training and test sets might have different genetic features affecting DoR. The model performances for RBN and SL were relatively low (< 0.2 and ~0, respectively, **Fig. 1b**, **Fig. S2a**). Variation in omics data and population structure only explains a small proportion of the trait variation. Therefore, variation in RBN and SL might be mainly explained by the environment and/or genotype-by-environment interactions.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To remove the potential confounding effects of kinship on mCor, we estimated linear regression models between mCor and kinship for training accessions (i.e., matrices containing only mCor and kinship values between all accessions (in both training and test sets) with accessions in the training set using the “lm” function in R, and then used the residuals of mCor to build a new Random Forest (RF) model.

Feature importance

Three measures were used to evaluate the importance or contributions of features to the prediction of complex traits: (1) RF “gini” importance, which reflects the impurity decrease when a feature is fed to the model[50]; (2) coefficients of features in the rrBLUP regression models; and (3) SHapley Additive exPlanations (SHAP) values, which reflect the contribution of a feature to the prediction of a complex trait[16]. The average absolute SHAP value of a feature for all the instances (local feature importance) was used to measure the global contribution of the feature to model prediction (global feature importance). The former two measures interpret global feature contributions to the model predictions and tend to assign non-zero importance values to all or most features (**Fig. 2**). In contrast, SHAP values provide local interpretations and tend to assign zero for features that have no contribution to the prediction of flowering time for individual instances (i.e., accessions in this study). A positive SHAP value indicates an instance is predicted to have higher trait value with a given feature than when that feature is removed from the model, and vice versa. The higher the absolute SHAP value, the more a feature contributes to trait prediction for the instance in question.

Feature selection

Because the number of predictors (1,771,290 G features, 24,175 T features, and 79,107 gbM features) was far greater than the number of observations (383 accessions) in this dataset, we wanted to prevent model overfitting. Thus, we conducted feature selection. The top 25, 50, 100, 200, 400, 800, 1600, 3200, 6400, 12,800, 25,600, 51,200, 102,400, 204,800, 409,600, 819,200, 1,638,400, and 3,276,800 features with highest importance values in RF models or absolute coefficients in rrBLUP models were selected to build new models. The optimal number of features was determined to be that leading to the highest model performance on CV sets. Feature selection led to limited improvement or even a decrease in the prediction performance (**Fig. S13,S14**), indicating that when all the features were used, there was no, or little, overfitting in our predictive models.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Benchmark flowering time genes

A list of 306 genes known to be involved in flowering in Arabidopsis (**Table S2**) were downloaded from the FLOR-ID database (http://www.phytosystems.ulg.ac.be/florid/)[17]. We also downloaded another 72 benchmark flowering time genes from the FLOR-ID database whose functions in flowering time are pending further experimental validation (referred to as 72 pending genes). To compare the importance of flowering genes across models built using different omics data, the highest importance (i.e., absolute coefficient in rrBLUP models, RF feature importance, and absolute SHAP values) of all SNPs (or methylation sites) within a gene or near a gene (the nearest gene) was used as the gene-based importance. When all three types of gene-body methylation data (i.e., CG, CGH, and CHH) were used together, the highest feature importance of the three types was used as the gene-based importance.

To understand why no significantly more benchmark genes were identified than random for most combinations of omics datasets and importance measures, we explored the following four potential reasons. First**,** we determined whether the threshold affects the number of identified benchmark genes by increasing the threshold of gene importance scores to the 99th percentile. The higher the threshold, the fewer benchmark genes were identified, but also the fewer genes were expected (1% for threshold at 99th percentile). We found that there were significantly more benchmark genes identified than random (1%) only when SHAP values from models built using transcriptome data were used to identify important genes (**Table S5**), indicating that SHAP values are able to reveal the most important genes (99th vs 95th percentile) for flowering time prediction. However, for all the other combinations of datasets and feature importance measures, no significant differences were observed between results when 95th and 99th percentile were used as thresholds, indicating that the choice of threshold only had a minor effect on the identification of benchmark genes.

Second**,** we asked whether inclusion of the 72 pending genes in the benchmark gene list affected the number of benchmark genes identified. We found that a relatively higher proportion of the 306 confirmed benchmark genes (i.e., those with experimental evidence) were identified compared with the 72 pending genes (**Table S4**, **Fig. S15**). However, when looking only at the 306 confirmed benchmark genes, no more genes were identified than expected randomly for most combinations of omics datasets and importance measures (**Table S4**).

Third**,** since the functions of benchmark genes in flowering may have been determined under different conditions (e.g., under a different temperature or photoperiod), we investigated

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

whether the use of flowering time data measured under only one condition (at 10 °C)[10] explained the failure to identify more benchmark genes than expected. We reasoned that benchmark genes showing effects on flowering under multiple conditions when mutated or overexpressed would be more likely to contribute to flowering at 10 °C than genes showing effects only under one condition. To test this hypothesis, we asked whether genes contributing to flowering under two different conditions (SDs and LDs, unfortunately no data for different temperatures) were more likely to be identified than those contributing to flowering only under one condition, but found that this was not the case (**Table S4**, **Fig. S15**).

Fourth, we asked whether the sample size of accessions (306 accessions in the training set, for which data for six traits and all the G, T, M data were available) was too small. To test this, we repeated the analysis using all 618 accessions (494 accessions were used to train the model, for which all the G, T, M data and flowering time information were available). Generally, no more benchmark genes were identified than when 306 accessions were used to train the models (**Table S6**), suggesting that decreasing the number of accessions included in the model from 618 to 383 was not a major factor affecting the identification of benchmark genes.

In summary, the four potential factors we considered only had minor, if any, effects on the number of known flowering genes identified in the models. Since no effect was identified, we continued with our original strategy, using the important genes with feature importance scores above the 95th percentile from models built using 306 accessions, for subsequent analysis.

In addition to the FLOR-ID database, we obtained 52 additional flowering genes by searching for the term “flowering time” in the TAIR database, and found that 24 out of these 52 genes were important for predicting flowering time in our models (**Table S10**). Consistent with the findings for FLOR-ID benchmark genes, G- and T-based RF models identified the highest number of genes (five and seven, respectively) from these 52 additional flowering genes (**Table S10**). In addition, there was no significant difference in the proportion of identified genes between FLOR-ID and the additional gene sets (odds ratio=0.75, p=0.37). Taken together, although our feature importance-based approaches didn’t identify significantly more flowering genes than random chance for most cases, our approaches outperformed the GWAS approach where only five QTLs were identified[10].

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Plant materials and assessment of flowering time

The potential function in flowering time of 21 genes with importance values above the 95th percentile (for at least one model based on one of the three omics data combined with one of three importance measures) was validated experimentally. These 21 important genes, together with another 37 non-important genes, were selected from an existing dataset in our lab to assess the degree of genetic redundancy between duplicate genes, and T-DNA insertion mutants of these genes in the Arabidopsis Col-0 background were used (**Table S12**). Generation and genotyping of mutants and the planting and growth conditions were described in detail previously[51]. In brief, seeds for homozygous mutant and WT sibling plants (using seeds collected from two to six plants per genotype; referred to as sublines) were planted to compare the flowering time between mutants and WT.

Days to flowering (number of days from placing flats containing seeds in the growth chamber until the appearance of the first open flower) were recorded for at least 40 plants per genotype (**Table S12**). Within each block, days to flowering observed for plants grown in different flats was first normalized using the function “normalize” of the R package “broman”. Average differences in flowering time between mutants and WT were calculated, and the significance of differences was assessed using the Wilcoxon rank sum test.

Feature integration

To investigate interactions among G, T, and gbM features of benchmark flowering time genes, we first built a model using all the features (G + T + gbM) related to the 378 benchmark genes. Then, to simplify the analysis, only one feature each for G (i.e., the SNP having the highest importance rank among other SNPs in a gene), T, and gbM (i.e., CG-, CHG-, or CHH-type gene-body methylation, having the highest importance rank among others in a gene) was kept for each gene. This resulted in 1084 features for the 378 genes. A new model was built using these top features, and the interactions among these features were calculated using the “TreeExplainer” and “shap_interaction_values” functions of the SHAP package[33]. The SHAP interaction of feature i and feature j can be interpreted as the difference in SHAP values of feature i with and without the feature j[33]. The higher the value, the stronger the interaction.

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Code availability

All the scripts and datasets used in this study are available on Github at:

https://github.com/ShiuLab/Manuscript_Code/tree/master/2023_Ath_multi-omics

Acknowledgements

We thank Elyse Vischulis for help in data collection, and Chad Niederhuth and Sunil Kumar Kenchanmane Raju for helpful discussions on DNA methylation. This work was supported by the U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409 to SHS), the National Science Foundation (DGE-1828149 and IOS-2218206 to SHS, KSA, and SL; IOS-2107215 and MCB-2210431 to MDL and SHS), and the Scientific Research Foundation of and the Major Scientific Research Tasks from Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003 and KIMA-ZDKY2022004 to PW).

Author Contributions

PW and SHS conceived and designed the study with input from SL, KSA, and MDL. PW and SL performed the methylation-related analysis. PW and KSA analyzed the feature importance. MDL conducted the assessment of flowering time for loss-of-function mutant plants. PW conducted all other analyses. PW, SL, KSA, MDL, and SHS wrote the paper. All authors read and approved the final manuscript.

References

1. VanRaden, P. M. *et al.* Invited Review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16–24 (2009).
2. Crossa, J. *et al.* Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science* **22**, 961–975 (2017).
3. Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G. & Shiu, S.-H. Transcriptome-Based Prediction of Complex Traits in Maize. *Plant Cell* **32**, 139–151 (2020).
4. Michel, S. *et al.* Merging genomics and transcriptomics for predicting fusarium head blight resistance in wheat. *Genes* **12**, 114 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5. Johannes, F. *et al.* Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**, e1000530 (2009).
6. Hu, Y., Morota, G., Rosa, G. J. M. & Gianola, D. Prediction of Plant Height in *Arabidopsis thaliana* Using DNA Methylation Data. *Genetics* **201**, 779–793 (2015).
7. Riedelsheimer, C. *et al.* Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* **44**, 217–220 (2012).
8. Hu, X., Xie, W., Wu, C. & Xu, S. A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnology Journal* **17**, 2011–2020 (2019).
9. Kawakatsu, T. *et al.* Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**, 492–505 (2016).
10. Alonso-Blanco, C. *et al.* 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
11. Grimm, D. G. *et al.* easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell* **29**, 5–19 (2017).
12. Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* **4**, 250–255 (2011).
13. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
14. Zhang, H., Lang, Z. & Zhu, J.-K. Dynamics and function of DNA methylation in plants. *Nat Rev Mol Cell Biol* **19**, 489–506 (2018).
15. Chien, P.-S., Chen, P.-H., Lee, C.-R. & Chiou, T.-J. TWAS coupled with eQTL analysis reveals the genetic connection between gene expression and flowering time in *Arabidopsis*. *Journal of Experimental Botany* erad262 (2023) doi:10.1093/jxb/erad262.
16. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv:1705.07874 [cs, stat]* (2017).
17. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res* **44**, D1167–D1171

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(2016).

18. Sheldon, C. C. *et al.* The *FLF* MADS Box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell* **11**, 445–458 (1999).
19. Williams, B. P., Bechen, L. L., Pohlmann, D. A. & Gehring, M. Somatic DNA demethylation generates tissue-specific methylation states and impacts flowering time. *The Plant Cell* **34**, 1189–1206 (2022).
20. Rosloski, S. M., Jali, S. S., Balasubramanian, S., Weigel, D. & Grbic, V. Natural diversity in flowering responses of *Arabidopsis thaliana* caused by variation in a tandem gene array. *Genetics* **186**, 263–276 (2010).
21. Yang, S. *et al.* Nitrilases NIT1/2/3 positively regulate flowering by inhibiting *MAF4* expression in *Arabidopsis*. *Front Plant Sci* **13**, 889460 (2022).
22. Liu, F. *et al.* The *Arabidopsis* RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate *FLC*. *Molecular Cell* **28**, 398–407 (2007).
23. Liu, F., Marquardt, S., Lister, C., Swiezewski, S. & Dean, C. Targeted 3′ processing of antisense transcripts triggers *Arabidopsis FLC* chromatin silencing. *Science* **327**, 94–97 (2010).
24. Zhang, L. & Jiménez□Gómez, J. M. Functional analysis of *FRIGIDA* using naturally occurring variation in *Arabidopsis thaliana*. *Plant J* **103**, 154–165 (2020).
25. Chiang, G. C. K. *et al.* Pleiotropy in the wild: the dormancy gene DOG1 exerts cascading control on life cycles. *Evolution* **67**, 883–893 (2013).
26. Ding, Z. J. *et al.* WRKY41 controls *Arabidopsis* seed dormancy via direct regulation of *ABI3* transcript levels not downstream of ABA. *Plant J* **79**, 810–823 (2014).
27. Cai, Y. *et al.* Dlf1, a WRKY transcription factor, is involved in the control of flowering time and plant height in rice. *PLoS ONE* **9**, e102529 (2014).
28. Azodi, C. B., Tang, J. & Shiu, S.-H. Opening the black box: interpretable machine learning for geneticists. *Trends Genet* **36**, 442–455 (2020).
29. Lee, J. H. *et al.* Integration of floral inductive signals by flowering locus T and suppressor of

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

overexpression of *Constans 1*. *Physiol Plant* **126**, 475–483 (2006).

30. Rouse, D. T., Sheldon, C. C., Bagnall, D. J., Peacock, W. J. & Dennis, E. S. FLC, a repressor of flowering, is regulated by genes in different inductive pathways: FLC protein levels and vernalization. *The Plant Journal* **29**, 183–191 (2002).
31. Moon, J., Lee, H., Kim, M. & Lee, I. Analysis of flowering pathway integrators in *Arabidopsis*. *Plant and Cell Physiology* **46**, 292–299 (2005).
32. Han, L. *et al.* A multi-omics integrative network map of maize. *Nat Genet* **55**, 144–153 (2023).
33. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. Preprint at https://doi.org/arXiv:1802.03888v3 (2019).
34. Aguirre, L., Hendelman, A., Hutton, S. F., McCandlish, D. M. & Lippman, Z. B. Idiosyncratic and dose-dependent epistasis drives variation in tomato fruit size. *Science* **382**, 315–320 (2023).
35. Deng, W. *et al.* FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6680–6685 (2011).
36. Wagner, D. Making flowers at the right time. *Developmental Cell* **37**, 208–210 (2016).
37. Fernandez, D. E. *et al.* The MADS-domain factors AGAMOUS-LIKE15 and AGAMOUS-LIKE18, along with SHORT VEGETATIVE PHASE and AGAMOUS-LIKE24, are necessary to block floral gene expression during the vegetative phase. *Plant Physiology* **165**, 1591–1603 (2014).
38. Hou, X. *et al.* Nuclear factor Y-mediated H3K27me3 demethylation of the *SOC1* locus orchestrates flowering responses of *Arabidopsis*. *Nat Commun* **5**, 4601 (2014).
39. Imura, Y. *et al.* CRYPTIC PRECOCIOUS/MED12 is a novel flowering regulator with multiple target steps in *Arabidopsis*. *Plant and Cell Physiology* **53**, 287–303 (2012).
40. Airoldi, C. A., McKay, M. & Davies, B. *MAF2* is regulated by temperature-dependent splicing and represses flowering at low temperatures in parallel with *FLM*. *PLoS ONE* **10**, e0126516 (2015).
41. He, Y. Chromatin regulation of flowering. *Trends in Plant Science* **17**, 556–562 (2012).
42. Bu, Z. *et al.* Regulation of *Arabidopsis* flowering by the histone mark readers MRG1/2 via

bioRxiv preprint doi: https://doi.org/10.1101/2023.11.14.566971; this version posted November 14, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

interaction with CONSTANS to modulate *FT* expression. *PLoS Genet* **10**, e1004617 (2014).

43. Tian, H. *et al.* Photoperiod-responsive changes in chromatin accessibility in phloem companion and epidermis cells of *Arabidopsis* leaves. *The Plant Cell* **33**, 475–491 (2021).
44. Zhao, B., Xi, Y., Kim, J. & Sung, S. Chromatin architectural proteins regulate flowering time by precluding gene looping. *Sci. Adv.* **7**, eabg3097 (2021).
45. Genger, R. K., Peacock, J. W., Dennis, E. S. & Finnegan, J. E. Opposing effects of reduced DNA methylation on flowering time in *Arabidopsis thaliana*. *Planta* **216**, 461–466 (2003).
46. Endelman, J. B. & Jannink, J.-L. Shrinkage estimation of the realized relationship matrix. *G3 (Bethesda)* **2**, 1405–1413 (2012).
47. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
48. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
49. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
50. Strobl, C., Boulesteix, A.-L. & Augustin, T. Unbiased split selection for classification trees based on the Gini Index. *Computational Statistics & Data Analysis* **52**, 483–501 (2007).
51. Wang, P. *et al.* High□throughput measurement of plant fitness traits with an object detection method using Faster R□CNN. *New Phytologist* **234**, 1521–1533 (2022).