PNAS

Supporting Information for

Diverse Genotype-by-Weather Interactions in Switchgrass

Alice H. MacQueen, Li Zhang, Samuel A. Smith, Jason Bonnette, Arvid R. Boe, Phillip A. Fay, Felix B. Fritschi, David B. Lowry, Robert B. Mitchell, Francis M. Rouquette Jr, Yanqi Wu, Arbel Harpak, and Thomas E. Juenger

Thomas E. Juenger
E-mail: tjuenger@utexas.edu

This PDF file includes:

Other supporting materials for this manuscript include the following:

Datasets S1 to S8

This PNAS template for Supporting Information (SI) may be used to organize your supporting material. The template is intended to provide a clearly organized PDF file that will ensure readers can easily navigate to sections or specific figures and tables. Movie files or large datasets can be presented as separate files. Further information is available in the PNAS Author Center.

Using the template. Specify the title, author list, and corresponding authors with the same YAML specification as in the corresponding PNAS Quarto template. The cover page will be automatically generated with the relevant description of the SI.

Figures should be placed on separate pages with legends set immediately below each figure. Table titles should be set immediately above each table. Note that tables extending beyond the width of the page can be included in the PDF or provided as separate dataset files. Oversized/nonstandard page sizes are accepted as part of your SI Appendix file.

Submitting SI. Save your completed SI file as a PDF for submission. Further submission instructions are available here.

Supplementary Methods

Section S1. Covariance matrices generated to provide multiple ways SNP effects could covary across common gardens. A trait measured in two environments can be considered two traits that are genetically correlated (Falconer 1952). By extension, a phenological trait measured at eight common gardens can be modeled as eight genetically correlated traits represented by a covariance matrix, where diagonal entries represent narrow-sense heritability at a common garden, or the additive genetic variance present in the population at that garden, and non-diagonal entries represent the genetic correlation between the trait in a pair of environments.

We were interested in the types of genetic correlation that could be expressed by different alleles across the genome and reasoned that different loci could have different patterns of genetic correlation. For example, the effects of a locus, A, could covary with a weather-based cue if A had a common allele, a, that was responsive to this cue. In contrast, the effects of a different locus B might have an effect at only one garden, for example if a pathogen was only present at one garden, and B has a common allele b that is resistant to that pathogen.

We specified three qualitative categories of covariance structures across gardens: "canonical" covariance, with simple patterns of effect size covariance introduced in the initial *mash* manuscript; "data-driven" covariances derived from common patterns of SNP effects observed in the data and introduced in the initial *mash* manuscript, and "GxWeather" covariance, estimated from the covariance of empirical weather patterns at each garden over specific time frames before each genotype underwent the phenological event.

Canonical covariance matrices were created by *mash* and fell into four groups: an identity matrix, an equal effects matrix, singleton matrices, and simple heterozygousity matrices. The identity matrix has values of zero for covariance between different gardens and correlations of one within gardens, and represents the situation in which the effects in different conditions are independent. The equal effects matrix has values of one for all matrix elements, and represents identical effects across all conditions. Each singleton matrix has a value of one for a single garden and zeros for all other elements, representing a garden-specific effect. Three simple heterozygousity matrices are generated, which have covariances of 0.25, 0.5, or 0.75 for all off-diagonal matrix elements, and correlations of one for diagonal entries, representing moderately positively correlated effects across all sites.

Data-driven covariance matrices can also be created by *mash* and fall into two groups: an overall pattern of covariance between all conditions, denoted as $tPCA$, and the first five principle components of a singular value decomposition of the $tPCA$, denoted as PCA_1 through PCA_5 (Urbut et al. 2018).

To create the GxWeather covariance matrices we transformed each genotype's green-up and flowering date at each garden to create date-related values that were summary functions for a weather variable across a specified date range. For example, the cumulative sum of the rainfall that occurred seven days prior to the flowering date was used to create one possible GxWeather cue variable for flowering date. The code for these calculations can be found at git@github.com:Alice-MacQueen/pvdiv-phenology-gxe/R/Analysis_v1.2_mash_using_greedy_covar.qmd, starting at line 390.

We defined the two phenological dates as functions of five weather variables and eight to ten time frames (Table 1). First, cumulative growing degree days (GDD) variables for each time frame were calculated as $GDD = \sum_{DATE-i}^{DATE} max(T_{mean} - T_{base}, 0)$, where T_{mean} is the daily average temperature, defined as $(T_{max} + T_{min})/2$, T_{base} is the base temperature of 12°C for switchgrass, T_{max} is the maximum daily temperature, T_{min} is the minimum daily temperature, $DATE$ is the phenological date, and i is the time frame, or the number of days prior to the phenological date that GDD is summed across (Kiniry et al. 2005; Behrman et al. 2013). Second, cumulative rainfall values in each time frame were calculated as $C_{rain} = \sum_{DATE-i}^{DATE} PRCP$, where PRCP is the daily rainfall measured in millimeters, $DATE$ is the phenological date, and i is the time frame, or the number of days prior to the phenological date that rainfall is summed across. Third, day length in hours was determined for each time frame as a specific single day prior to the phenological date, as indicated by the time frame i, e.g. $DATE - i$. Day length was calculated as a function of latitude and day of the year as in (Forsythe et al. 1995). Fourth, day length change in seconds was determined for a specific single day prior to the phenological date as indicated by the time frame, as $DYLN_{change(s)} = (DYLN_{DATE-i} - DYLN_{DATE-i-1}) * 60 * 60$, where $DYLN$ in the day length in hours, $DATE$ is the

phenological date, and i is the number of days prior to the phenological date. Fifth, average temperature in each time frame was defined as $T_{ave} = \sum_{DATE-i}^{DATE}(T_{max} + T_{min})/2$, where T_{max} is the maximum daily temperature, T_{min} is the minimum daily temperature, $DATE$ is the phenological date, and i is the time frame, or the number of days prior to the phenological date that average temperature is summed across.

We then generated GxWeather covariance matrices derived from correlations in these GxWeather cue variables for the three subpopulations. These covariance matrices represent the correlations between genotypes for these weather-related variables across our eight common gardens. To create these matrices, we determined the correlations between these weather-related cue variables for identical genotypes grown at different common gardens and used these correlations to fill the off-diagonals of a eight-by-eight matrix. To fill the diagonal entries of these matrices, we used the narrow-sense heritability for each weather-related cue variable at each garden, calculated using the same methods as Section S3. Matrices need to be positive semi-definite to be used in *mash*, so when the resultant matrices were not positive semi-definite, the `Matrix::nearPD` function was used to compute the nearest positive definite matrix to the correlation matrix (Bates, Maechler, and Jagan 2023). Finally, all covariance matrices were rescaled so that the maximum diagonal element of the rescaled matrix is 1, as in (Urbut et al. 2018). Code used to generate these matrices can be found at git@github.com:Alice-MacQueen/pvdiv-phenology-gxe.git/R/Analysis_v1.2_mash_using_greedy_covar.qmd, starting at line 716.

We defined distinct sets of GxWeather covariance matrices for the two phenological dates and for the three genetic subgroups. We used the same five weather variables to transform both green-up and flowering date, but defined the weather-based covariance matrices relative to each of the two phenological dates. GxWeather covariance matrices were also defined separately for individual genotypes from the Gulf, Midwest, and Both subpopulations.

Section S2. Univariate genome-wide association to obtain initial effect estimates. To obtain initial effect estimates for each single nucleotide polymorphism (SNP) for each combination of phenological trait, genetic subpopulation, and common garden, we conducted univarate genome-wide association (GWAS) on site-specific best linear unbiased predictors (BLUPs) using the switchgrassGWAS package and the methods in (Lovell et al. 2021). The full SNP diversity panel generated in (Lovell et al. 2021) was subset within each subpopulation for all subsequenct analyses to retain only SNPs with $\leq 20\%$ missing data and minor allele frequencies > 0.05. This subsetting resulted in 8.8 million SNPs retained for the Midwest subpopulation, 10.3 million SNPs retained for the Gulf subpopulation, and 12.3 million SNPs retained for Both subpopulations.

BLUPs were calculated independently for each combination of trait, genetic subpopulation, and common garden. We used ASReml to estimate BLUPs using mixed models of the form:

$$\mathbf{y} = 1 + Zu + e$$

$$Var(u) = G\sigma_u^2$$

$$Var(e) = I\sigma_e^2$$

where the vector $\mathbf{y}$ represents the flowering date or greenup date values for that garden, Z the design matrix for random effects, u the whole genome additive genetic effect, e the residual, G the genomic relationship matrix based on all SNPs retained for subpopulation-specific analyses, and I the rank-$\mathbf{y}$ identity matrix (Clark and Werf 2013). ASReml can specify considerably more complex variance structures; however, we were interested in specifying covariance structures using *mash*, which allows the specification of multiple covariance structures; hence, we used a relatively simple model to calculate BLUPs for use in univariate GWAS. The code for these calculations can be found at git@github.com:Alice-MacQueen/pvdiv-phenology-gxe.git/R/Analysis_v1.2_mash_using_greedy_covar.qmd, starting at line 74.

Univariate GWAS were run independently for each combination of trait, genetic subpopulation, and common garden. The code used to run GWAS can be found at git@github.com:Alice-MacQueen/pvdiv-phenology-gxe.git/R/Analysis_v1.2_mash_using_greedy_covar.qm starting at line 240. We used `pvdiv_standard_run()`, an opinionated function of the switchgrassGWAS R package, to conduct GWAS. To control for population structure and run GWAS, this function uses a three-part process. First, a fast truncated singular value decomposition (SVD) is used, which has initial pruning of regions in long-range linkage disequilibrium (LD) and iterative pruining and removal of regions in long-range LD (Privé, Aschard, and Blum 2017a). By default, 15 vectors of singular values, k, are computed. Pruning and removal of long-range LD regions are recommended in human genetics to best control for population structure (Abdellaoui et al. 2013; Privé, Aschard, and Blum 2017b; Price et al. 2008). Second, linear regression models using 0-15 of these vectors of singular values were used to run GWAS and to calculate the genomic control coefficient, λ_{GC}, for each k (Devlin and Roeder 1999). Third, we chose the lowest value of k that made λ_{GC} closest to 1 to correct for population structure, and ran GWAS using a linear regression model with this k for population structure correction.

Section S3. Selection of effect estimate subsets to model using *mash*. To make the subsequent analyses using *mash* computationally feasible, we followed the author's recommended analysis outline for large datasets using *mash* and considered subsets of the millions of effects at different stages of the mash analysis. We extracted two subsets of tests from our complete GWAS effect estimate datasets: 1) effects from a subset of "strong" tests corresponding to stronger effects on our traits; 2) results from a random subset of all tests to correspond to an unbiased representation of all effects. We also clumped effects before selection so that signal strength was standardized across different LD blocks in the genome, and not limited to closely linked associations.

To prepare subsets of effect estimates and their standard errors from these GWAS for use in mash, we used `pvdiv_bigsnp2mashr()` from the switchgrassGWAS R package. For each phenological trait and each genetic subpopulation, we created two subsets of effects and standard errors in paired $m \times n$ matrices, where m were marker effects and n were common gardens. The first subset of effects and standard errors, the strong effects, was created by selecting relatively unlinked ($r^2 < 0.2$) SNPs from each univariate GWAS that had the smallest p-values in those GWAS. Notably, the majority of these "strong" effects were not significant at a 5% level in univariate GWAS. The number of SNPs from each univariate GWAS was chosen to result in approximately million entry matrices for each trait & subpopulation. We selected 19K - 33K effects and standard errors for the strong effect subsets (Table S1). In many cases, there was overlap in the markers with the largest effect estimates at different locations; thus, 19 - 33K SNPs selected for each of eight univariate GWAS ultimately gave sets of 35K - 63K SNPs (Table S1). The second subset of effects and standard errors, the random subset, was created by selecting relatively unlinked ($r^2 < 0.2$) SNPs from each univariate GWAS after clumping using minor allele frequency. To account for the chance that true effects could be sparse, we doubled the size of the random subset relative to the strong subset. We selected 38 - 66K SNP effects and formed subsets of 71 - 127K SNP markers for the random subset.

Section S4. Greedy *mash* algorithm used to select covariance matrices that significantly improved the model log-likelihood. Because we were particularly interested in determining whether genomic loci varied in their patterns of genetic covariance across gardens, specifying the optimal set of covariance matrices in *mash* was an important model-building problem. If very similar covariance matrices are included when fitting a model in *mash*, *mash* will arbitrarily assign effect loadings across these matrices, as all assignments will lead to the same fit to the effect-size data. Our GxWeather covariance matrices were in many cases highly correlated, as they represented covariation in the same weather variables in overlapping time frames. We thus evaluated the performance of *mash* models when different covariance matrices were included by implementing a "greedy" mash algorithm.

Section S5. *Mash* to jointly re-estimate SNP effects across eight common gardens.

Section S6. Narrow-sense heritability estimation. In the diversity panel, we determined narrow-sense heritabilities (h^2) for green-up and flowering dates at single gardens using genomic relationship matrices calculated using the van Raden method (VanRaden 2008). Genomic relationship matrices were calculated within each subpopulation (Midwest and Gulf) and for both genetic subpopulations (Both). We used rrBLUP (Endelman 2011) to specify mixed models of the form:

$$\mathbf{y} = 1 + Zu + e$$

$$Var(u) = G\sigma_u^2$$

$$Var(e) = I\sigma_e^2$$

in which the vector $\mathbf{y}$ represents the flowering date or green-up date values for that garden, Z the design matrix for random effects, u the whole genome additive genetic effect, and e the residual. Matrix G is the whole genomic relationship matrix based on all SNPs retained for subpopulation-specific analyses. I is the rank-y identity matrix. Phenotypic variance σ_p^2 is $\sigma_u^2 + \sigma_e^2$. Narrow-sense heritability is then $h^2 = (\sigma_u^2/\sigma_p^2)$.

These models were run for each of the eight gardens, and across all gardens by adding an additional environmental effect of site without an interaction term. This resulted in 54 models: 3 sets of populations (the Gulf, Midwest, and Both subpopulations) for 9 garden sets (all eight gardens separately, and all eight gardens together) and two phenotypes (green-up date and flowering date).

Section S7. Outbred pseudo-F2 mapping population and Quantitative Trait Locus mapping. To confirm candidate genomic regions and patterns of allelic effects found in the diversity panel, we analyzed flowering in an outbred pseudo-F2 cross between four grandparents, two Midwest and two Gulf individuals. The formation of this mapping population has been described previously (Milano, Lowry, and Juenger 2016). The parents of this cross were DAC, an early flowering Midwest individual, VS16, a late flowering Midwest individual, AP13, an early flowering Gulf individual, and WBC, a late flowering Gulf individual. We made F1 crosses of the two early flowering genotypes, AP13xDAC, and the two late flowering genotypes, WBCxVS16. We then clonally propagated and planted the four parents, the two F1 genotypes (AP13xDAC, and VS16xWBC), and 801 F2

genotypes at eight field sites in May-July of 2015. To be directly comparable to the diversity panel data, only 2019 phenology data from the pseudo-F2 cross from the same eight common garden sites were used here.

Details on the genetic map construction, map polishing and fine-scale reordering can be accessed on DataDryad. QTL mapping was conducted with R/qtl2 (Broman et al. 2019). We performed a genome scan with a linear mixed model that accounts for the relationships among individuals and for environmental covariates (i.e., field sites). The full model can be expressed as:

$$phenotype = \mu + QTL + E + QTL * E + kinship + e$$

where μ is the population mean, QTL is the marker genetic effect, E is the environmental effects (here, common garden), QTL*E is the interaction between marker genetic and environmental effects, kinship corresponds to the background polygenic variation, and e is the error term. The genome scan was accomplished with the 'scan1' function. The statistical significance of the genome scan was established by performing a stratified (i.e., stratifying on common garden) permutation test (n=1000) using 'scan1perm' function. The estimated QTL effect was obtained using 'scan1coef' function in R/qtl2.

Figures should be cross-referenced like Figure S1. Each figure should be on its own page. You can control this by placing a pagebreak shortcode, with .

This can also be done with LaTeX's `\newpage` command. You may find `\FloatBarrier` useful if your figures or tables are running off to the wrong pages.

Figure S1. Figure

Table S1. Effect estimates selected for joint re-estimation across eight common gardens using *mash*

Phenology type	Subpopulation	Effects Selected per GWAS, Strong Subset	Effects Selected per GWAS, Random Subset	Total Markers, Strong Subset	Total Markers, Random Subset
Greenup Date	Gulf	19000	38000	45931	91862
	Midwest	24000	48000	49030	98060
	Both	19000	38000	44765	89530
Flowering Date	Gulf	19000	38000	50490	100980
	Midwest	33000	66000	63367	126734
	Both	19000	38000	35735	71470

Datasets. If your document relies on movies or datasets, please list them here with their captions. Use the `movie{your caption}` and `\dataset{file_name.ext}{your caption}` commands to do so.

Datasets 1 through 6 are csv files where the first two columns have the following definitions:

- Marker: The SNP marker in the format Chromosome_Position
- log10BF: log10(Bayes Factor) of the significance of the marker effect in the mash model

The remaining column names follow the pattern Effect_[Mean/StandardError/lfsr]_[Subpopulation]_[Phenotype]_[Garden], where Mean and Standard Error are estimates of the effect mean and standard error, lfsr is the local false sign rate statistic for the effect, and [Subpopulation], [Phenotype], and [Garden] follow the conventions of Figure 1.

SI Dataset S1 (Dataset_1_Gulf_vegetative_growth_date.csv)

SNP-associated effects for the start of vegetative growth jointly re-estimated in the Gulf genetic subpopulation.

SI Dataset S2 (Dataset_2_Midwest_vegetative_growth_date.csv)

SNP-associated effects and standard errors for the start of vegetative growth jointly re-estimated in the Midwest genetic subpopulation.

SI Dataset S3 (Dataset_3_Both_subpopulations_vegetative_growth_date.csv)

SNP-associated effects and standard errors for the start of vegetative growth jointly re-estimated in both the Midwest and Gulf genetic subpopulations.

SI Dataset S4 (Dataset_4_Gulf_reproductive_growth_date.csv)

SNP-associated effects and standard errors for the start of reproductive growth jointly re-estimated in the Gulf genetic subpopulation.

SI Dataset S5 (Dataset_5_Midwest_reproductive_growth_date.csv)

SNP-associated effects and standard errors for the start of reproductive growth jointly re-estimated in the Midwest genetic subpopulation.

SI Dataset S6 (Dataset_6_Both_subpopulations_reproductive_growth_date.csv)

SNP-associated effects and standard errors for the start of reproductive growth jointly re-estimated in both the Midwest and Gulf genetic subpopulations.

SI Dataset S7 (Dataset_7_QTL_intervals_with_functionally_validated_homologs.csv)

Genes in the quantitative trait loci (QTL) regions shown in Figure 4 that have functionally validated homologs in rice for vegetative or reproductive development. The first 9 columns of this table provide QTL information, the remainder provide annotation information for the switchgrass genes in the QTL interval with homologs in A. thaliana or rice that have functional validation in flowering or vegetative-growth related traits. Columns 20-30 provide the titles of papers providing functional validation of the rice homolog of the switchgrass gene.

SI Dataset S8 (Dataset_8_QTL_mash_overlap.csv)

Overlap between quantitative trait loci (QTL) for the start of vegetative or reproductive growth with significant mash effect estimates (with a log10(Bayes Factor) > 2) found using diversity panels of the Midwest, Gulf, and Both genetic subpopulations. The first five columns of this table provide QTL information, and columns 6-10 provide mash Marker information for markers within the QTL regions.

References

Abdellaoui, Abdel, Jouke-Jan Hottenga, Peter de Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, et al. 2013. "Population Structure, Migration, and Diversifying Selection in the Netherlands." *European Journal of Human Genetics* 21 (11): 1277–85. https://doi.org/10.1038/ejhg.2013.48.

Bates, Douglas, Martin Maechler, and Mikael Jagan. 2023. "Matrix: Sparse and Dense Matrix Classes and Methods." https://CRAN.R-project.org/package=Matrix.

Behrman, Kathrine D., James R. Kiniry, Michael Winchell, Thomas E. Juenger, and Timothy H. Keitt. 2013. "Spatial Forecasting of Switchgrass Productivity Under Current and Future Climate Change Scenarios." *Ecological Applications* 23 (1): 73–85. https://doi.org/10.1890/12-0436.1.

Broman, Karl W, Daniel M Gatti, Petr Simecek, Nicholas A Furlotte, Pjotr Prins, Śaunak Sen, Brian S Yandell, and Gary A Churchill. 2019. "R/Qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations." *Genetics* 211 (2): 495–502. https://doi.org/10.1534/genetics.118.301595.

Clark, Samuel A., and Julius van der Werf. 2013. "Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values." In, 321–30. Humana Press. https://doi.org/10.1007/978-1-62703-447-0_13.

Devlin, B., and Kathryn Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55 (4): 997–1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x.

Endelman, J. B. 2011. "Ridge Regression and Other Kernels for Genomic Selection with r Package rrBLUP" 4: 250–55.

Falconer, D. S. 1952. "The Problem of Environment and Selection." *The American Naturalist* 86 (830): 293–98. https://doi.org/10.1086/281736.

Forsythe, William C., Edward J. Rykiel, Randal S. Stahl, Hsin-i Wu, and Robert M. Schoolfield. 1995. "A Model Comparison for Daylength as a Function of Latitude and Day of Year." *Ecological Modelling* 80 (1): 87–95. https://doi.org/10.1016/0304-3800(94)00034-f.

Kiniry, J. R., K. A. Cassida, M. A. Hussey, J. P. Muir, W. R. Ocumpaugh, J. C. Read, R. L. Reed, M. A. Sanderson, B. C. Venuto, and J. R. Williams. 2005. "Switchgrass Simulation by the ALMANAC Model at Diverse Sites in the Southern US." *Biomass and Bioenergy* 29 (6): 419–25. https://doi.org/10.1016/j.biombioe.2005.06.003.

Lovell, John T., Alice H. MacQueen, Sujan Mamidi, Jason Bonnette, Jerry Jenkins, Joseph D. Napier, Avinash Sreedasyam, et al. 2021. "Genomic Mechanisms of Climate Adaptation in Polyploid Bioenergy Switchgrass." *Nature* 590 (7846): 438–44. https://doi.org/10.1038/s41586-020-03127-1.

Milano, Elizabeth R, David B Lowry, and Thomas E Juenger. 2016. "The Genetic Basis of Upland/Lowland Ecotype Divergence in Switchgrass (*Panicum Virgatum*)." *G3 Genes|Genomes|Genetics* 6 (11): 3561–70. https://doi.org/10.1534/g3.116.032763.

Price, Alkes L., Michael E. Weale, Nick Patterson, Simon R. Myers, Anna C. Need, Kevin V. Shianna, Dongliang Ge, et al. 2008. "Long-Range LD Can Confound Genome Scans in Admixed Populations." *The American Journal of Human Genetics* 83 (1): 132–35. https://doi.org/10.1016/j.ajhg.2008.06.005.

Privé, Florian, Hugues Aschard, and Michael G. B. Blum. 2017b. "Efficient Management and Analysis of Large-Scale Genome-Wide Data with Two r Packages: Bigstatsr and Bigsnpr." http://dx.doi.org/10.1101/190926.

———. 2017a. "Efficient Management and Analysis of Large-Scale Genome-Wide Data with Two r Packages: Bigstatsr and Bigsnpr." http://dx.doi.org/10.1101/190926.

Urbut, Sarah M., Gao Wang, Peter Carbonetto, and Matthew Stephens. 2018. "Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions." *Nature Genetics* 51 (1): 187–95. https://doi.org/10.1038/s41588-018-0268-8.

VanRaden, P. M. 2008. "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science* 91 (11): 4414–23. https://doi.org/10.3168/jds.2007-0980.