University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Biochemistry -- Faculty Publications

Biochemistry, Department of

2019

Genome-wide association and genomic prediction for biomass yield in a genetically diverse *Miscanthus sinensis* germplasm panel phenotyped at five locations in Asia and North America

Lindsay V. Clark
University of Illinois at Urbana-Champaign

Maria S. Dwiyanti
Hokkaido University

Kossonou G. Anzoua
Hokkaido University

Joe E. Brummer
Colorado State University

Bimal Kumar Ghimire
Konkuk University

See next page for additional authors

Clark, Lindsay V.; Dwiyanti, Maria S.; Anzoua, Kossonou G.; Brummer, Joe E.; Ghimire, Bimal Kumar; Glowacka, Katarzyna; Hall, Megan; Heo, Kweon; Jin, Xiaoli; Lipka, Alexander E.; Peng, Junhua; Yamada, Toshihiko; Yoo, Ji Hye; Yu, Chang Yeon; Zhao, Hua; Long, Stephen P.; and Sacks, Erik J., "Genome-wide association and genomic prediction for biomass yield in a genetically diverse *Miscanthus sinensis* germplasm panel phenotyped at five locations in Asia and North America" (2019). *Biochemistry -- Faculty Publications*. 426.
https://digitalcommons.unl.edu/biochemfacpub/426

Authors
Lindsay V. Clark, Maria S. Dwiyanti, Kossonou G. Anzoua, Joe E. Brummer, Bimal Kumar Ghimire, Katarzyna Glowacka, Megan Hall, Kweon Heo, Xiaoli Jin, Alexander E. Lipka, Junhua Peng, Toshihiko Yamada, Ji Hye Yoo, Chang Yeon Yu, Hua Zhao, Stephen P. Long, and Erik J. Sacks

Received: 20 December 2018 | Revised: 24 March 2019 | Accepted: 26 March 2019

DOI: 10.1111/gcbb.12620

ORIGINAL RESEARCH

WILEY GLOBAL CHANGE BIOLOGY BIOENERGY

Genome-wide association and genomic prediction for biomass yield in a genetically diverse *Miscanthus sinensis* germplasm panel phenotyped at five locations in Asia and North America

Lindsay V. Clark[1] | **Maria S. Dwiyanti**[2] | **Kossonou G. Anzoua**[2] | **Joe E. Brummer**[3] | **Bimal Kumar Ghimire**[4] | **Katarzyna Głowacka**[5] | **Megan Hall**[6] | **Kweon Heo**[7] | **Xiaoli Jin**[8] | **Alexander E. Lipka**[1] | **Junhua Peng**[9] | **Toshihiko Yamada**[2] | **Ji Hye Yoo**[7] | **Chang Yeon Yu**[7] | **Hua Zhao**[10] | **Stephen P. Long**[1] | **Erik J. Sacks**[1]

[1]Department of Crop Sciences, University of Illinois, Urbana-Champaign, Urbana, Illinois

[2]Field Science Center for Northern Biosphere, Hokkaido University, Sapporo, Japan

[3]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado

[4]Department of Applied Bioscience, Konkuk University, Seoul, South Korea

[5]Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska

[6]Bio Architecture Lab, Berkeley, California

[7]Department of Applied Plant Sciences, Kangwon National University, Chuncheon, Gangwon, South Korea

[8]Department of Agronomy, Zhejiang University, Hangzhou, China

[9]HuaZhi Biotechnology Institute, Changsha, China

[10]College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China

Correspondence
Erik J. Sacks, Department of Crop Sciences, University of Illinois, Urbana-Champaign, 1201 W. Gregory Dr., Urbana, IL 61801, USA.
Email: esacks@illinois.edu

Funding information
Office of Science; DOE Office of Science, Office of Biological and Environmental Research, Grant/Award Number: DE-SC0006634 DE-SC0018420

Abstract

To improve the efficiency of breeding of *Miscanthus* for biomass yield, there is a need to develop genomics-assisted selection for this long-lived perennial crop by relating genotype to phenotype and breeding value across a broad range of environments. We present the first genome-wide association (GWA) and genomic prediction study of *Miscanthus* that utilizes multilocation phenotypic data. A panel of 568 *Miscanthus sinensis* accessions was genotyped with 46,177 single nucleotide polymorphisms (SNPs) and evaluated at one subtropical and five temperate locations over 3 years for biomass yield and 14 yield-component traits. GWA and genomic prediction were performed separately for different years of data in order to assess reproducibility. The analyses were also performed for individual field trial locations, as well as combined phenotypic data across groups of locations. GWA analyses identified 27 significant SNPs for yield, and a total of 504 associations across 298 unique SNPs across all traits, sites, and years. For yield, the greatest number of significant SNPs was identified by combining phenotypic data across all six locations. For some of the other yield-component traits, greater numbers of significant SNPs were obtained from single site data, although the number of significant SNPs varied greatly from site to site. Candidate genes were identified. Accounting for population structure, genomic prediction accuracies for biomass yield ranged from 0.31 to 0.35 across five northern sites and from 0.13 to 0.18 for the subtropical location, depending on the estimation method. Genomic prediction accuracies of all traits were similar for single-location and multilocation data, suggesting that genomic selection will be useful for breeding broadly adapted *M. sinensis* as well as *M. sinensis* optimized for specific climates. All of our data, including DNA sequences flanking each SNP, are publicly available. By facilitating genomic selection in *M. sinensis* and *Miscanthus* × *giganteus*, our results will accelerate the breeding of these species for biomass in diverse environments.

KEYWORDS

biomass yield, field trials, genome-wide association studies, genomic selection, *Miscanthus sinensis*, *Miscanthus* × *giganteus*, RAD-seq

1 | INTRODUCTION

Miscanthus is a promising crop for lignocellulosic bioenergy and bioproducts, but it is in the very early stages of domestication and breeding (Clifton-Brown, Chiang, & Hodkinson, 2008; Clifton-Brown et al., 2019; Dwiyanti, Stewart, & Yamada, 2013; Sacks, Juvik, Lin, Stewart, & Yamada, 2013). A single sterile triploid clone of *Miscanthus* × *giganteus*, a hybrid between *Miscanthus sacchariflorus* and *Miscanthus sinensis*, has been adopted for commercial biomass production in Europe and North America (Głowacka et al., 2015; Heaton, Clifton-Brown, Voigt, Jones, & Long, 2004). However, new biomass cultivars of *Miscanthus* are needed to limit the risk of disease and pest outbreaks associated with growing a monoculture of a single clone. Additionally, the standard clone of *M.* × *giganteus* is insufficiently winter-hardy in parts of northern Europe and northern parts of the US Midwest (Clifton-Brown & Lewandowski, 2000; Dong et al., 2019), and it flowers too early to yield optimally at lower latitudes (~30°) such as the coastal plain of the southern United States (E. J. Sacks, unpublished data). Both *M. sinensis* and *M. sacchariflorus* are native to a broad range of environments across East Asia and should provide a wealth of breeding material for developing new biomass cultivars (Clifton-Brown et al., 2008). Such breeding can be accelerated by genomic selection, especially because *Miscanthus* is a long-lived perennial that requires 3 years of field testing to obtain high-quality yield data. Additionally, candidate genes and associated single nucleotide polymorphisms (SNPs) identified by genome-wide association (GWA) can be used to enhance genomic prediction (Bian & Holland, 2017; Rice & Lipka, 2019; Spindel et al., 2016), and identify targets for genome editing (Scheben & Edwards, 2018). By obtaining knowledge about associations between genotype and biomass yield and 14 yield-component traits, for a diverse germplasm panel of *M. sinensis* evaluated at one subtropical and five temperate locations, we aim to build a strong foundation for genomic-enabled breeding of this new crop.

Previous GWA and genomic prediction studies of *Miscanthus* have used phenotypic data collected from single locations (Nie et al., 2016; Slavov et al., 2014; Zhao et al., 2013). Slavov et al. (2014) conducted a GWA study and genomic prediction on 138 genotypes of *M. sinensis* from the Korean Peninsula and Southern Japan (S Japan), using more than 100,000 SNP markers. Slavov et al. (2014) phenotyped the *M. sinensis* accessions for yield traits near Aberystwyth, United Kingdom, and found significant SNP–trait associations for stem and leaf length, stem angle, senescence, and lignin content, but none for overall dry matter yield. Dry matter yield in the Slavov et al. (2014) study and a follow-up study by Davey et al. (2017) had moderate broad-sense heritability but low genomic predictive ability (0.04–0.06), which they hypothesized could be improved considerably if the study included a larger number of individuals. It was also demonstrated that genomic selection could be more successful if a selection index were used in place of direct selection for yield (Davey et al., 2017; Slavov et al., 2019). Zhao et al. (2013) studied 300 *M. sinensis* genotypes collected from a broad geographic range across China in a field trial at Wuhan, China, using 115 alleles from 23 microsatellite markers. They identified nine significant associations with heading date, plant height, and yield, although a relatedness matrix was not included in the GWA model, increasing the chance of false positives. Nie et al. (2016) used a unified mixed linear model (MLM; Yu et al., 2006) with 1,059 markers from several polymerase chain reaction (PCR)-based genotyping methods on a collection of 138 *M. sinensis* genotypes from southwest China phenotyped in a field trial in Sichuan and identified one significant association with biomass yield and 11 significant associations with other traits. Genomic prediction accuracies estimated by Nie et al. (2016) were generally lower than those estimated by Slavov et al. (2014), although the estimate for dry biomass yield was higher (0.23).

Given the high genetic diversity, population structure (Clark et al., 2014) and known genotype-by-environment effects (Arnoult & Brancourt-Hulmel, 2015; Clifton-Brown et al., 2001; Kaiser, Clark, Juvik, Voigt, & Sacks, 2015; Yan et al., 2012) in *M. sinensis*, inferences from GWA and genomic prediction would benefit from multilocation phenotypic data on large, genetically diverse germplasm panels. GWA and genomic prediction studies that used phenotypic data spanning multiple continents are rare, with notable exceptions being a few studies in wheat (Crossa et al., 2014, 2007). However, such broad sampling of environments can enable the identification of SNPs that are broadly useful for improving breeding values across environments (Wei et al., 2010) and improve genomic prediction accuracy via the use of phenotypes from correlated environments (Crossa et al., 2014; Spindel et al., 2016).

In this study, we present the first multilocation GWA and genomic prediction study in *M. sinensis*. In total, 568 genotypes previously characterized for population structure (Clark et al., 2014) were phenotyped at six field trial locations in Asia and North America (Clark et al., 2019), and were genotyped with 46,177 SNPs using RAD-seq. Our goals were to (a) identify candidate genes for yield and yield-component traits; (b) assess reproducibility of GWA results across years, locations, and correlated traits; and (c) estimate genomic prediction accuracy in order to determine the feasibility of genomic selection.

2 | MATERIALS AND METHODS

2.1 | Plant material and phenotypic data collection

Dry biomass yield and 14 yield-component traits (Table 1) were recorded for 568 *M. sinensis* clonal genotypes at six field trial locations in the second and third year after planting as previously described (Clark et al., 2019). In short, field trials were established early in the 2012 growing season at five temperate locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); and Chuncheon, Korea by Kangwon National University (KNU); and at one southern (subtropical) location, Zhuji, China by Zhejiang University (ZJU). Field trials were randomized complete block designs with three to four replicate at each site; plots were single plants equally spaced within and between rows on 1.5 m centers. Harvesting was conducted in late autumn or early winter, after dormancy or the first killing freeze led to dry down, with stems being cut 15–20 cm above the ground. *M. sinensis* genotypes included wild accessions representing six genetic groups in East Asia, plus ornamental and US naturalized accessions that clustered with an S Japan based on marker data (Clark

TABLE 1 Yield and yield-component traits measured in multilocation field trials of *Miscanthus sinensis*

Trait	Abbreviation	Description[a] and notes
Dry biomass yield (g/plant)	Yld	Single plant plots on 1.5 m centers were harvested in late autumn by cutting the stems 15–20 cm above the soil surface. Samples were dried at 60°C until constant weight. Estimates are reported per area based on plot dimensions (2.25 m^2) or per plant
Compressed circumference (cm)	CC	Stems were compressed at the middle height of the plant such that all the culms were in close contact without air gaps; then the circumference of the compressed bundle was measured
Basal circumference (cm)	BC	Circumference of the base of the plant, without compression
Compressed circumference/basal circumference	CC/BC	Compressed circumference divided by basal circumference, to estimate the proportion of the plant's footprint filled by stems
Culm length (cm)	CmL	Length of the tallest culm in late autumn, measured from the base of the stem to the tip of the panicle if present, otherwise to the highest part of the highest leaf, following the standard evaluation system for measuring height in rice (International Rice Research Institute, 2002)
Culm node number	CmNdN	Number of nodes on the tallest culm of each plant in late autumn
Internode length (cm)	IntL	Culm length divided by the number of nodes for the tallest culm of each plant in late autumn
Culm dry weight (g)	CmDW	Mass of the tallest culm of each plant in late autumn, after removal of leaves and drying at 60°C until constant weight achieved (not recorded at KNU in year 3 or 4 or at ZJU)
Culm volume (cm^3)	CmV	Estimated from culm length, culm diameter at first internode, and culm diameter at last internode, assuming the stem was shaped like the frustum of a cone: $\mathrm{CmL} \times \pi \times [(\mathrm{DBI}/2)^2 + (\mathrm{DBI}/2) \times (\mathrm{DTI}/2) + (\mathrm{DTI}/2)^2]/3$
Culm density (g/cm^3)	CmDW/V	Culm dry weight divided by culm volume (not estimated for KNU in year 3 or 4 or at ZJU)
Diameter of basal internode (mm)	DBI	Measured on the tallest culm of each plant in late autumn
Diameter of topmost internode (mm)	DTI	Measured on the tallest culm of each plant in late autumn
Total number of culms	TCmN	Counted for each plant
Proportion of reproductive culms	RCmN/TCmN	Number of reproductive culms divided by the total number of culms (not estimated at HU in year 2 or at KNU)
Culms per footprint (#/cm^2)	TCmN/A	The total number of culms divided by the area of the plant's footprint. The footprint area was estimated from the basal circumference, assuming a circular base

Abbreviations: KNU, Kangwon National University; ZJU, Zhejiang University.

[a] All traits were measured at the end of the growing season.

et al., 2014). Pairwise Jost's *D* (Jost, 2008) among the eight genetic groups ranged from 0.01 to 0.08, indicating low genetic differentiation consistent with *M. sinensis* being a highly outcrossing species (Clark et al., 2014).

2.2 | Quantitative genetics

Best linear unbiased predictor (BLUP) values were calculated for subsequent use in GWA and genomic prediction. Phenotypic data were transformed using the Box–Cox method (Box & Cox, 1964) prior to BLUP calculation, as previously described (Clark et al., 2019), in order to prevent false positive associations, inflated prediction accuracy estimates, and other issues that can arise from violation of model assumptions in GWA and genomic prediction (Owens et al., 2014). Random effects models were fitted using the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015). Models were fit for each of 188 location × trait × year combinations (Equation 1), with replication (*R*) and genotype (*G*) as random effects. In Equation 1, *Y* is the vector of Box–Cox-transformed phenotypes, *b* is the intercept, β is a vector of coefficients (β_2 being the BLUPs used in downstream analysis), and ε is random error.

$$Y=b+\beta_1R+\beta_2G+\varepsilon \tag{1}$$

Additionally, 132 multilocation models (Equation 2, Table 2) were fit using the site combinations HU + NEF, ZJU + KNU, HU + NEF + UI + CSU, HU + NEF + UI + CSU + KNU (all five northern trial sites), and HU + NEF + UI + CSU + KNU + ZJU (all six trial sites), which were chosen based on environmental similarities, genetic correlations, and ANOVA of phenotypic values among locations (Clark et al., 2019), in order to generate multilocation BLUPs for GWA and genomic prediction. Multilocation models were only fitted when data for a given trait × year combination were available for all locations in a given combination of locations. Location (*L*), replication within location, genotype, and genotype × location were included as random effects in the multilocation models.

$$Y=b+\beta_1R(L)+\beta_2G+\beta_3GL+\beta_4L+\varepsilon \tag{2}$$

However, if Equations 1 and 2 are applied to all entries in the germplasm panel, they do not account for population structure. Prior studies have found that population structure can bias estimates of genomic prediction accuracies upwards (Fiedler et al., 2018; Guo et al., 2014; Riedelsheimer et al., 2012). Because we previously found that the genetic groups within *M. sinensis* differed from each other with respect to yield and yield-component traits, and because it was possible to predict from marker data which genetic group a genotype was in, we expected genomic prediction accuracy across the entire germplasm collection to be higher than genomic prediction accuracy within groups. Estimation of prediction accuracy within genetic groups would therefore serve as an indicator of how well genomic selection would work if one were to breed within groups, or within a population where individuals from different genetic groups were intermated a sufficient number of generations to greatly reduce or eliminate population structure. Genetic groups in the *M. sinensis* diversity panel included S Japan (which included ornamental and US naturalized accessions), Northern Japan (N Japan), Korea/Northern China (N China), Yangtze-Qinling, Sichuan, Southeast China (SE China), as determined by the discriminant analysis of principal components (DAPC; Clark et al., 2014).

We evaluated three methods for estimating genomic prediction accuracy within genetic groups: (a) performing genomic prediction separately within each genetic group using BLUPs from Equations 1 and 2, (b) using genotype-within-genetic group BLUPs (Equations 3 and 4, below) in genomic prediction, and (c) fitting the BLUPs from Equations 1 and 2 to the genetic group and using the residuals in genomic prediction (Equation 5, below). Method (a) closely resembles the genomic prediction that might be performed by plant breeders wishing to select within individual genetic groups, but in this study, estimates from this strategy were limited by small sample sizes within genetic groups. In contrast, methods (b) and (c) allowed us to more accurately estimate the prediction accuracy without bias from population structure from the representative panel that we studied. Method (b) was designed to integrate coherently with our method for estimating BLUPs, and to control for interaction effects between genetic group and location. Method (c) was based on a method that Lipka et al. (2014) developed previously to control for population structure in genomic prediction. All three methods were performed using traits measured in year 3 only.

For method (a) of controlling for population structure, genomic prediction was run within genetic group for groups where at least 50 genotypes had phenotypic data at a given site or site combination, using BLUPs from Equations 1 and 2. The US naturalized and ornamental accessions were included as part of S Japan for this purpose, given their known S Japan ancestry.

For method (b) of controlling for population structure, genotypic BLUPs were calculated where between-group variance was removed, leaving only within-group variance. BLUPs for genotype-within-genetic-group within each individual location were calculated with Equation 3, where *D* is the genetic group.

$$Y=b+\beta_1R+\beta_2D+\beta_3G(D)+\varepsilon \tag{3}$$

Multilocation BLUPs for genotype-within-genetic-group were calculated with Equation 4 for all five northern trial sites (HU + NEF + UI + CSU + KNU), and for all six trial sites (HU + NEF + UI + CSU + KNU + ZJU).

TABLE 2 Number of significant SNP–trait associations detected in genome-wide association analyses of *Miscanthus sinensis*. Phenotypic values were Box–Cox transformed, then genotype BLUPs were calculated for each trait. In total, 46,177 SNP markers and 568 accessions were included in genome-wide association analyses. Q + K mixed model analyses were performed using the software GAPIT. Associations were considered significant if $p < 0.05$ after FDR correction

Trait	Year	HU	NEF	ZJU	HU + NEF	HU + NEF + CSU + UI	HU + NEF + CSU + UI + KNU	HU + NEF + CSU + UI + KNU + ZJU	KNU + ZJU	Total[a]
Dry biomass yield (g/plant)	2	0	0	3	0	0	0	0	2	27
	3	0	0	0	0	0	0	26	3	
	4	0	NM	NM	NC	NC	NC	NC	NC	
Compressed circumference (cm)	2	0	0	2	0	0	0	0	0	9
	3	0	6	1	0	0	0	0	0	
Basal circumference (cm)	2	0	3	SK	1	1	1	0	SK	7
	3	0	SK	SK	SK	SK	SK	1	2	
Compressed circumference/basal circumference	2	0	5	1	0	0	0	0	0	30
	3	7	2	15	2	1	0	0	0	
Culm length (cm)	2	0	21	0	0	0	0	0	ZH	24
	3	0	1	1	0	0	0	1	0	
Culm node number	2	0	0	2	1	1	1	2	ZH	11
	3	4	3	0	1	1	1	0	0	
Internode length (cm)	2	3	3	23	7	4	1	2	9	43
	3	0	2	0	0	0	0	0	0	
Culm dry weight (g)	2	0	0	NM	1	1	0	NC	NC	1
	3	0	0	NM	0	0	NC	NC	NC	
Culm volume (cm^3)	2	1	0	0	0	0	0	0	0	1
	3	0	0	0	0	0	0	0	0	
Culm density (g/cm^3)	2	25	2	NM	52	42	37	NC	NC	65
	3	4	0	NM	2	3	NC	NC	NC	
Diameter of basal internode (mm)	2	1	2	0	2	2	6	1	5	12
	3	3	0	0	1	1	5	5	0	
Diameter of topmost internode (mm)	2	1	0	0	0	0	0	0	27	28
	3	0	0	0	0	0	0	0	0	
Total number of culms	2	0	0	0	0	0	0	0	0	48
	3	47	0	0	14	3	1	0	0	
Proportion of reproductive culms	2	NM	0	0	NC	NC	NC	NC	NC	8
	3	0	8	0	1	1	NC	NC	NC	
Culms per footprint ($\#/cm^2$)	2	0	0	0	0	0	0	0	0	12
	3	4	0	1	2	2	2	1	5	
Total[b]		100	58	49	87	63	55	39	53	326

Abbreviations: SNP, single nucleotide polymorphism; BLUP, best linear unbiased predictor; HU, Hokkaido University; NEF, New Energy Farms; CSU, Colorado State University; UI, University of Illinois at Urbana-Champaign; KNU, Kangwon National University; ZJU, Zhejiang University; NM, trait was not measured; NC, multisite genotypic BLUP not calculated because trait was not measured at one or more sites; SK, skewness ≥2 or ≤−2 led us to exclude the genotypic BLUP from the GWAS; ZH, zero heritability across sites, therefore not calculated.

[a] Number of SNPs with any significant associations across all years, sites, and site combinations, within traits.

[b] Total across all traits.

$$Y = b + \beta_1 L + \beta_2 R(L) + \beta_3 D + \beta_4 DL + \beta_5 G(D) + \beta_6 G(D)L + \varepsilon \tag{4}$$

The BLUP values used in downstream genomic prediction analysis from Equation 3 was represented by β_3, and for Equation 4, β_5 was the BLUP used.

For method (c) of controlling for population structure, models were fitted with genetic group as a fixed effect and BLUPs from Equations 1 and 2 as the response variable. Residuals from these models were then used as the response variable in genomic prediction (see below). These residuals were calculated for traits measured in year 3 only. Residuals were calculated with Equation 5, where D is the genetic group.

$$\beta_2 = b + \beta_5 D + \varepsilon \tag{5}$$

To improve genomic estimated breeding values by utilizing genetic correlations among traits, selection indices were estimated for each individual location, the five northern locations, and all six locations, using an approach similar to that of Davey et al. (2017). Year 3 measurements of Yld, CC, BC, CmL, CmNdN, DBI, and TCmN were included in the selection index. CmDW, CmDW/V, and TCmN/RCmN were excluded as these were not measured at all sites. To avoid error from the inclusion of highly correlated variables (Baker, 1986), CC/BC, IntL, CmV, and TCmN/A were excluded as these were calculated from other traits in the model, and DTI was excluded for its strong correlation with DBI. Plots that did not have year 3 measurements for all seven traits were excluded. For each location and location combination, phenotypic (***P***) and genetic (***G***) variance–covariance matrices were estimated using Box–Cox transformed data. The genetic variance was estimated from Equation 1 (individual locations) or Equation 2 (location combinations). The genetic covariance was estimated as $(\sigma^2_{G(\text{trait1 + trait2})} - \sigma^2_{G(\text{trait1})} - \sigma^2_{G(\text{trait2})})/2$, where $\sigma^2_{G(\text{trait1})}$ and $\sigma^2_{G(\text{trait2})}$ are the genetic variances of individual traits, and $\sigma^2_{G(\text{trait1 + trait2})}$ is the genetic variance of the sum of the two traits. The economic value vector ***a*** was set to 1 for Yld and 0 for all other traits. The weighting factor w was then estimated as:

$$\boldsymbol{w} = \boldsymbol{P}^{-1}\boldsymbol{Ga} \tag{6}$$

The selection index ***S*** was then estimated as.

$$\boldsymbol{S} = \boldsymbol{Xw} \tag{7}$$

where ***X*** is a matrix of phenotypic observations by plot. Equations 1–4 were then used to estimate BLUPs of the selection index, and Equation 5 was used to estimate residuals of regressing BLUPs on DAPC groups.

2.3 | Genotyping

RAD-seq was performed according to a previously described protocol (Clark et al., 2014). In summary, genomic DNA was digested with *Msp*I and either *Pst*I-HF or *Nsi*I-HF (New England BioLabs). Digested DNA was then ligated to a barcoded adapter with a *Pst*I/*Nsi*I overhang and a universal adapter with an *Msp*I overhang. Ninety-five barcoded samples were then pooled into one library, and a 200–500 bp size selection was performed on 2% agarose. Libraries were amplified with a Kapa Hi-Fi library amplification kit (Kapa Biosystems, Wilmington, Massachusetts, USA), quantified, and then sequenced on a HiSeq 2500 (Illumina) at the Roy J. Carver Biotechnology Center at the University of Illinois. Nine *Pst*I libraries from a previous study (Clark et al., 2014) as well as eight additional *Pst*I libraries and thirteen *Nsi*I libraries (data available at NCBI Sequence Read Archive, accession SRP026347), were included in the analysis. Every individual in the study was represented on at least two *Pst*I libraries and two *Nsi*I libraries.

SNPs were mined from RAD-seq data on 594 accessions (Data S1), including 568 *M. sinensis* accessions with phenotypic data, 3 doubled haploid *M. sinensis* genotypes that were not included in the field trials, 3 *M. sinensis* accessions that did not survive at any field trial location, 9 diploid and 1 triploid *M.* × *giganteus* accession, and 7 diploid and 3 tetraploid *M. sacchariflorus* accessions. SNPs were called with the UNEAK pipeline (Lu et al., 2013) using a minimum call rate of 0.04 and a minimum minor allele frequency of 0.002 for initial filtering. Additionally, 402 GoldenGate SNPs from a previous study (Clark et al., 2014) were included. SNPs were then imported into R, and any SNPs that appeared heterozygous in any of the doubled haploid lines were removed from the dataset because these were evidence of allelic differences at paralogous loci in the recently duplicated diploid *M. sinensis* genome. The dataset was then filtered to only include SNPs that had missing data in 30% or fewer individuals and a minor allele frequency of at least 0.05 in at least one of the nine genetic groups (eight *M. sinensis* groups and one *M. sacchariflorus* group) previously identified by DAPC (Clark et al., 2014; Jombart, Devillard, & Balloux, 2010). A total of 46,177 SNP markers (including GoldenGate SNPs) were retained, with an overall missing data rate of 26% averaged across all accessions and SNPs. After removing 20 individuals that were *M. sacchariflorus* or F_1 *M.* × *giganteus* hybrids, imputation of missing SNPs was performed for the remaining 574 *M. sinensis* individuals (including the three doubled haploid lines), using the estimation–maximization method based on relatedness (Poland et al., 2012) as implemented in the R package rrBLUP (Endelman, 2011).

To obtain genomic positions of SNPs for identification of candidate genes, sequence tags for all SNPs were aligned to the *Sorghum bicolor* 3.0 reference genome (DOE-JGI, http://phytozome.jgi.doe.gov/) using Bowtie2 (Langmead & Salzberg, 2012) under relaxed parameters (*-D 20 -R 3 -N 1 -L 18 -i S,1,0.50 --local*). Alignment positions were obtained for 26,804 SNPs. Alignments were queried against the *S. bicolor* 3.1 genome annotation, revealing that 20,611 SNPs were within 1 kb of at least one protein-coding gene.

2.4 | Genome-wide association

Unified mixed linear model genome-wide association (MLM GWA) analyses were performed on 568 *M. sinensis* individuals (all *M. sinensis* with phenotype and genotype data; Data S1) using 46,177 imputed SNP markers. MLM GWA was implemented with the software GAPIT (Lipka et al., 2012) version 2016.03.01 using kinship matrix compression and P3D (Zhang et al., 2010) under default parameters. The kinship matrix included in the model was estimated by GAPIT using the method of VanRaden (2008). Included as covariates were seven columns of Q values estimated by Structure (Falush, Stephens, & Pritchard, 2003) from our previous study of *M. sinensis* population structure (Clark et al., 2015; Data S1). The eighth column of Q values was omitted since it could be predicted from the other seven. The eight columns of Q values corresponded to ancestry from *M. sinensis* populations in S Japan, central Japan, N Japan, Korea/N China, Sichuan, Yangtze-Qinling, and SE China/tropical, as well as *M. sacchariflorus*. The S Japan population from Clark et al. (2014) corresponded to the S Japan and central Japan populations from Clark et al. (2015), due to larger sample size in the latter study resulting in finer resolution of population structure in Japan. All single-site BLUPs and multisite BLUPs were individually subjected to GWA. Upon examination of the number of significant SNPs identified, GWA results were only retained for further analysis if the trait BLUP included values for more than 200 genotypes and if its skewness was no less than −2 and no greater than +2, so that associations would not be driven by accessions with extreme phenotypes. All single-site BLUPs for CSU, UI, and KNU were excluded because each had fewer than 200 genotypes with data. Data from different years were treated separately in order to help assess the reproducibility of associations. In total, 83 single-site trait BLUPs and 127 multisite trait BLUPs were analyzed. For any given BLUP, SNPs were excluded from GWA if there were fewer than three genotypes with phenotypic data that also had the minor allele. Associations were considered significant if false discovery rate (FDR)-corrected *p*-values (Benjamini & Hochberg, 1995) were below 0.05.

2.5 | Genomic prediction

Genomic prediction for each location and for combinations of locations was performed using ridge regression-best linear unbiased prediction (RR-BLUP; Meuwissen, Hayes, & Goddard, 2001). Assuming equal phenotypic variance explained at each marker across the genome, an additive relationship matrix calculated from imputed SNP markers was used as the independent variable in the genomic prediction model, rather than the matrix of marker values (Endelman, 2011). The relationship matrix was calculated using the *A.mat* function, and the genomic prediction model was then calculated using the *kin.blup* function in the rrBLUP package (Endelman, 2011). Response variables for the genomic prediction models were the BLUP values obtained from phenotype data or the residuals of those BLUPs fitted to a model with genetic group as a fixed effect (Equations 1–5). Tenfold cross-validation was used as described by Resende et al. (2012). Briefly, for each trait, all genotypes having phenotype data were divided into 10 subgroups of equal size. Nine subgroups were used as a training set, while the remaining subgroup was used as a prediction set. The process was repeated 10 times in order to predict the genomic estimated breeding value (GEBV) once for each genotype. Prediction accuracy was estimated as the Pearson's correlation coefficient between the observed BLUP values and the predicted values. The stability of the model was evaluated by performing 100 iterations of cross-validation and estimating the average value of the Pearson's correlation coefficient from each iteration.

All data and code are available in the Illinois Data Bank, https://doi.org/10.13012/B2IDB-0790815_V3.

3 | RESULTS

3.1 | Genome-wide association

Across the 210 trait-site-year combinations examined, a total of 504 significant SNP–trait associations were identified at 5% FDR (Table 2). Within most traits, some SNPs were significant for more than one site, site combination, and/or year, yielding 326 significant associations when these duplicates were removed (Table 2). Because some SNPs were significant for more than one trait, 298 unique significant SNPs were identified (Data S2). Of all the significant associations, 307 were from traits measured in year 2 and 197 were from traits measured in year 3 (Table 2), and 16 SNPs had significant associations with at least one trait in both years 2 and 3 (Table S1). Most sets of traits that shared significant associations with a SNP were related to each other, such as culm diameter and culms per area (Table S1). An average of 2.5 significant SNPs were found for each trait × site × year estimated from single-site genotypic BLUPs, and 2.3 were found per trait × site combination × year based on multisite genotype BLUPs. The greater the number of sites used to calculate the multisite BLUPs, the fewer significant SNPs that were typically identified, with 3.1 and 2.4 for HU + NEF and KNU + ZJU respectively, 2.3 for HU + NEF + CSU + UI, 2.2 for HU + NEF + CSU + UI + KNU, and 1.6 for HU + NEF + CSU + UI + KNU + ZJU (Table 2). However, for yield, only 3 significant year 3 yield SNPs were identified for ZJU + KNU, whereas 26 were found for HU + NEF + UI + KNU + ZJU, (Figure 1; Table 2). In total, 27 unique SNPs for yield were identified, with 26 in year 3, 1 in year 2, and 2 in both years. The minor allele was associated with higher yield for 23 out of the 27 significant SNPs (Data S2). Eighteen of the significant SNPs for yield could be aligned to *S. bicolor* chromosomes, and the remaining nine were unaligned (Data S2). Of the yield SNPs aligned to the *S. bicolor* genome, five were on chromosome 1, three on chromosome 2, three on chromosome 3, two on chromosome 4, one on chromosome 6, one on chromosome 8, two on chromosome 9, and one on chromosome 10, with the closest pair of yield SNPs being over 1 Mb apart on chromosome 9 (Data S2, Figure 1 and Figure S1).

3.2 | Genomic prediction

Using genotypic BLUPs (Equations 1 and 2), genomic prediction accuracies for year 3 were mostly moderate for dry biomass yield and compressed circumference, the component trait that best predicted yield (Table 3). At each site, genomic prediction accuracies for culm length were higher than they were for yield or compressed circumference. KNU had a relatively small number of surviving genotypes and low replication (Clark et al., 2019), which likely contributed to its low prediction accuracies (Table 3; Table S2). Prediction accuracies for multisite genotype BLUPs (Equation 2) were comparable to those for single-site genotype BLUPs (Equation 1), suggesting that genomic selection can be used to identify genotypes that are high yielding across a broad range of environments (Table 3). Predicted breeding values are provided in Data S1. For all traits, prediction accuracies ranged from −0.19 to 0.79 for single-site BLUPs and 0.17–0.65 for multisite BLUPs (Table S2). Genomic prediction accuracies for diameter of basal internode, culm length, and culm volume were high (≥0.5) at all locations except at KNU, which had moderate values (Table 3; Table S2).

Genomic prediction accuracies for genotypes within genetic groups (using BLUPs calculated according to Equations 3 and 4 and residuals calculated according to Equation 5) were typically lower than those that did not control for population structure, but remained moderate in magnitude for most traits (Table 3), suggesting potential for genomic selection within genetic groups. Some notable exceptions were observed in which prediction accuracy was reduced drastically by accounting for population structure, including basal circumference at NEF and ZJU; compressed circumference/basal circumference at NEF and UI; diameter of topmost internode at HU; total number of culms at KNU; culms per footprint at NEF and ZJU; and yield, culm length, total number of culms, and culms per footprint at ZJU. On average, prediction accuracies were similar using the BLUP method (Equations 3 and 4) and the residuals method (Equation 5), although there were notable differences. For example, with respect to the method that did not control for population structure, prediction accuracies for basal circumference across multiple sites was reduced by ~50% when using residuals, but was reduced to zero when using genotype-within-genetic group BLUPs (Table 3). When yield BLUPs across temperate locations were used without controlling for the effect of genetic group (Equation 2), genomic prediction ranked the genotypes somewhat according to their genetic group (5.6% of variance in yield GEBVs explained by genetic group vs. 2.2% of variance in yield BLUPs explained by genetic group), suggesting that population structure played a role in prediction accuracy (Figure 2a). However, when the effect of genetic group was removed by any of our three methods (prediction within groups, within group BLUPs, or residuals), genomic prediction no longer ranked genotypes according to genetic group but was still moderately accurate (Figure 2b–d).

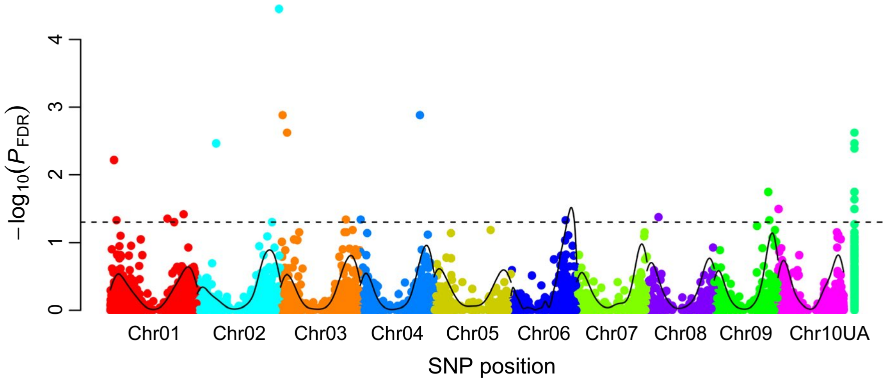


FIGURE 1 Manhattan plot indicating significance of SNP–trait associations for year 3 dry biomass yield identified in *Miscanthus sinensis* using genotypic best linear unbiased predictor (BLUPs) calculated across six field trial locations. The y-axis indicates log-transformed FDR-corrected *p*-values. The dashed line indicates the significance threshold at 5% FDR (26 significant single nucleotide polymorphisms [SNPs] shown). In total, 46,177 SNP markers and 568 accessions were included in genome-wide association analysis. Positions of SNPs with respect to the *Sorghum bicolor* v. 3.0 reference genome are indicated, and SNPs that were not aligned to the reference are placed in the rightmost group (UA). Solid curves indicate overall SNP density

TABLE 3 Genomic prediction accuracies for dry biomass yield and yield-component traits measured in year 3 on *Miscanthus sinensis* genotypes grown at five northern field trial locations (HU, NEF, CSU, UI, and KNU; 37.9–43.1°N) or one southern location (ZJU; 29.8°N). Phenotypic values were Box–Cox transformed, then genotype BLUPs (*G*; Equations 1 and 2), genotype-within-genetic group BLUPs (*G*(*D*); Equations 3 and 4), or residuals of genotype BLUPs fitted to genetic group (*R*; Equation 5) were calculated for each trait and used as the response variable in genomic prediction. *G*(*D*) BLUPs and *R* residuals eliminated phenotypic variance attributable to differences among genetic groups, in order to estimate the efficacy of genomic selection within genetic groups. The analyses were based on 46,177 SNP markers and 568 accessions

	BLUP type	HU	NEF	CSU	UI	KNU	ZJU	Northern trial locations	All trial locations
Selection index for dry biomass yield (Equation 7)	*G*	0.50	0.49	0.75	0.51	0.18	0.69	0.60	0.64
	G(D)	0.40	0.35	0.73	0.49	0.18	0.14	0.50	0.49
	R	0.38	0.33	0.71	0.48	0.12	0.09	0.49	0.47
Dry biomass yield (g/plant)	*G*	0.39	0.42	0.64	0.38	0.06	0.65	0.47	0.49
	G(D)	0.31	0.28	0.61	0.35	−0.01	0.13	0.35	0.17
	R	0.27	0.16	0.59	0.32	−0.08	0.18	0.31	0.29
Compressed circumference (cm)	*G*	0.35	0.37	0.58	0.41	0.36	0.50	0.45	0.44
	G(D)	0.23	0.22	0.56	0.32	0.33	0.20	0.37	0.32
	R	0.20	0.13	0.52	0.31	0.29	0.13	0.34	0.33
Basal circumference (cm)	*G*	0.37	0.34	0.52	0.56	0.23	0.24	0.44	0.39
	G(D)	0.17	−0.10	0.52	0.53	0.26	−0.15	−0.01	0.00
	R	0.12	−0.06	0.46	0.53	−0.01	−0.12	0.18	0.16
Compressed circumference/basal circumference	*G*	0.41	0.46	−0.06	0.28	0.27	0.53	0.48	0.51
	G(D)	0.22	0.04	0.02	−0.06	0.25	0.21	0.26	0.26
	R	0.16	0.02	−0.03	0.09	0.08	0.15	0.23	0.25
Culm length (cm)	*G*	0.52	0.56	0.79	0.61	0.46	0.72	0.61	0.54
	G(D)	0.36	0.44	0.72	0.52	0.40	−0.09	0.46	0.43
	R	0.35	0.44	0.71	0.53	0.37	−0.12	0.45	0.41
Culm node number	*G*	0.50	0.54	0.49	0.54	0.41	0.55	0.58	0.58
	G(D)	0.45	0.45	0.44	0.50	0.40	0.22	0.52	0.47
	R	0.45	0.45	0.43	0.50	0.41	0.23	0.52	0.49
Internode length (cm)	*G*	0.43	0.61	0.58	0.58	0.30	0.44	0.60	0.55
	G(D)	0.37	0.54	0.32	0.57	0.19	0.12	0.54	0.51
	R	0.35	0.53	0.33	0.55	0.15	0.01	0.54	0.50
Culm dry weight (g)	*G*	0.59	0.56	0.34	0.63				
	G(D)	0.41	0.33	0.18	0.52				
	R	0.44	0.38	0.17	0.57				
Culm volume (cm^3)	*G*	0.53	0.57	0.68	0.61	0.42	0.72	0.61	0.60
	G(D)	0.32	0.35	0.68	0.47	0.22	0.21	0.37	0.35
	R	0.36	0.41	0.66	0.53	0.32	0.23	0.44	0.41
Culm density (g/cm^3)	*G*	0.52	0.69	0.11	0.40				
	G(D)	0.28	0.46	−0.03	0.38				
	R	0.25	0.44	−0.03	0.37				
Diameter of basal internode (mm)	*G*	0.57	0.56	0.66	0.61	0.36	0.65	0.62	0.65
	G(D)	0.42	0.44	0.63	0.56	0.23	0.26	0.46	0.46
	R	0.41	0.45	0.61	0.56	0.25	0.24	0.48	0.47
Diameter of topmost internode (mm)	*G*	0.48	0.54	0.57	0.65	0.34	0.63	0.59	0.48
	G(D)	0.07	0.27	0.56	0.54	0.20	0.36	0.32	0.29
	R	0.00	0.26	0.55	0.55	0.08	0.35	0.32	0.21

(Continues)

TABLE 3 (Continued)

	BLUP type	HU	NEF	CSU	UI	KNU	ZJU	Northern trial locations	All trial locations
Total number of culms	*G*	0.39	0.40	0.60	0.52	0.27	0.38	0.50	0.48
	G(D)	0.33	0.24	0.43	0.33	0.04	−0.10	0.34	0.33
	R	0.09	0.26	0.40	0.37	−0.14	−0.09	0.24	0.23
Proportion of reproductive culms	*G*	0.44	0.55	0.57	0.61		0.70		
	G(D)	0.15	0.36	0.55	0.54		0.22		
	R	0.04	0.34	0.46	0.53		0.24		
Culms per footprint (#/cm^2)	*G*	0.56	0.54	0.28	0.62	0.31	0.39	0.60	0.60
	G(D)	0.57	−0.18	0.18	0.07	0.13	−0.07	0.26	0.26
	R	0.44	0.04	0.29	0.60	0.22	−0.09	0.40	0.40

Abbreviations: SNP, single nucleotide polymorphism; BLUP, best linear unbiased predictor; HU, Hokkaido University; NEF, New Energy Farms; CSU, Colorado State University; UI, University of Illinois at Urbana-Champaign; KNU, Kangwon National University; ZJU, Zhejiang University.

Among the genetic groups, S Japan (including ornamental and US naturalized accessions, known to be derived from S Japan germplasm) had the highest prediction accuracies overall, which were consistently moderate to high across locations and traits (Tables 4 and 5). The Yangtze-Qinling group had consistently moderate prediction accuracies among locations for compressed circumference/basal circumference, culm length, culm node number, and internode length, whereas other traits had moderate prediction accuracies at some locations and low prediction accuracies at others for this group (Table 4). N Japan had mostly moderate prediction accuracies at NEF, but moderate to low prediction accuracies at HU (Table 4). Low to moderate accuracies were observed in the Korea/N China group, but no traits were consistently moderate across locations in this group, despite this group having more genotypes than any other (Table 4). The SE China/tropical group only had enough genotypes for genomic prediction at ZJU, where prediction accuracies were mostly moderate (Table 4).

The use of a selection index for yield, based on biomass yield plus six yield-component traits, generally gave higher prediction accuracies than those observed for yield alone (Tables 3–5). The improvement in prediction accuracy was particularly consistent for multilocation data, with or without controlling for population structure. For example, prediction accuracy across all field trial locations without controlling for population structure was 0.64 for the selection index versus 0.49 for yield, and using genotype-within-genetic-group BLUPs was 0.49 for the selection index versus 0.17 for yield (Table 3). Due to environmental differences and different subsets of genotypes surviving at different locations, weighting factors for the selection index varied from location to location, although the weighting factor for total number of culms was always negative (Table S3), consistent with highly tillering plants tending to have short, thin culms that resulted in low yields (Clark et al., 2019).

4 | DISCUSSION

4.1 | SNP–trait associations and genomic predictions

Our results indicate that GWAs and genomic selection are likely to accelerate the breeding of *M. sinensis*, even with relatively low coverage of the genome by RAD-seq SNPs. Marker-assisted breeding via genomic selection can be expected to reduce the duration of the breeding cycle (seed to seed) of *M. sinensis* from approximately 4 years for conventional breeding based solely on phenotypic selection, to one to one and a half years for purely marker-assisted breeding (Clifton-Brown et al., 2008; Głowacka, 2011). Combinations of marker-assisted breeding and phenotypic selection are likely to be needed to confirm and revise genomic predictions, and such strategies might have breeding cycles of intermediate duration, although it could also be possible to maintain a minimum breeding cycle duration while concurrently phenotyping a subset of progeny of each generation for long-term model improvement. In addition to their potential use as covariates for genomic selection, the SNPs that we identified via GWA can be used to identify candidate genes to help elucidate the genetic architecture of complex traits in the grass family (Liu & Yan, 2019). In particular, given that *Miscanthus* is undomesticated and has a broad geographic distribution, we expect that the degree and pattern of genetic variation across its genome is very different from that of cereal crops and other domesticated grasses, affecting which biologically relevant genes are detectable in GWA. Thus, GWA of *Miscanthus* can help complement existing knowledge of phenotypic pathways in the grasses.

Using only 46,177 RAD-seq SNPs, we identified hundreds of significant associations with yield-related traits, including 34 for biomass yield per se across 27 unique SNPs. Most SNP–trait associations that we identified were not significant in more than 1 year, but those that were

(Table S1) represent high priority targets for additional study. Although for most traits SNP–trait associations based on single-location BLUPs were more frequent than those based on multilocation BLUPs, for biomass yield the greatest number of associations were detected using BLUPs that were calculated across all sites, despite a large genotype × environment effect on yield (Table 2; Clark et al., 2019). Due to differential survival of genotypes across field trial locations, the BLUPs calculated across all locations included more genotypes than for subsets of locations, which likely contributed greater statistical power to detect SNPs associated with yield. These associations based on multilocation BLUPs represent allelic effects that were consistent for trial locations, which we would expect to be advantageous for breeding cultivars with broad adaptation.

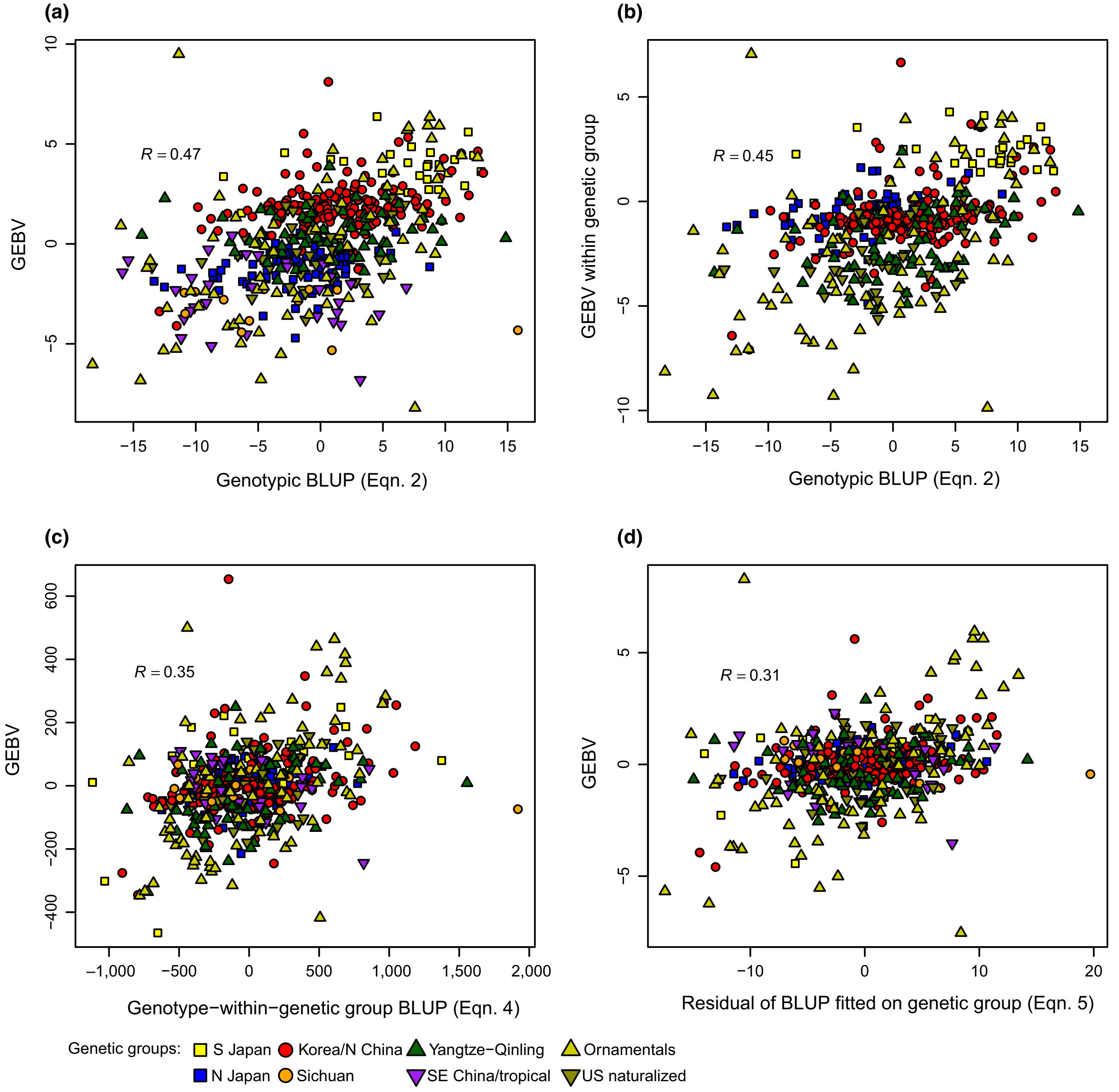


FIGURE 2 Genomic estimated breeding values (GEBVs) for year 3 dry biomass yield in *Miscanthus sinensis* across five temperate locations versus values used in genomic prediction. Colors indicate the genetic group to which each genotype belongs. (a) Genomic prediction using genotypic best linear unbiased predictor (BLUP) values without accounting for the genetic group (Equation 2). (b) Genomic prediction within genetic groups using BLUP values from Equation 2. (c) Genomic prediction using genotype-within-genetic-group BLUPs (Equation 4). (d) Genomic prediction using residuals of BLUP values regressed on the genetic group (Equation 5)

TABLE 4 Genomic prediction accuracies within genetic groups for traits measured in *Miscanthus sinensis* in year 3 at four northern field trial locations (HU, NEF, UI, and KNU; 37.9–43.1 °N) or one southern location (ZJU; 29.8 °N). Phenotypic values were Box–Cox transformed, then genotypic BLUPs were calculated (Equation 1) and used as the response variable in genomic prediction. The analyses were based on 46,177 SNP markers and 568 accessions. Genetic groups were excluded from analysis when fewer than 50 genotypes had phenotypic data

	HU				NEF				UI	KNU		ZJU		
	S Japan	N Japan	Korea/N China	Yangtze-Qinling	S Japan	N Japan	Korea/N China	Yangtze-Qinling	S Japan	S Japan	Korea/N China	Korea/N China	Yangtze-Qinling	SE China/ tropical
$N_{genotypes}$	117	75	151	51	137	83	156	64	84	73	56	77	63	66
Selection index (Equation 7)	0.69	0.15	0.05		0.53	0.42	0.41	0.39	0.55	0.39	−0.19	−0.42	0.54	0.33
Dry biomass yield (g/plant)	0.55	−0.08	0.11		0.43	0.34	0.38	0.25	0.42	0.19	−0.13	−0.43	0.51	0.28
Compressed circumference (cm)	0.54	−0.16	0.04	0.33	0.43	0.34	0.24	0.39	0.45	0.43	0.37	−0.29	0.56	0.23
Basal circumference (cm)	0.51	−0.08	−0.14	0.05	0.46	0.51	0.12	0.19	0.60	0.42	0.15	−0.56	0.38	0.18
Compressed circumference/basal circumference	0.46	−0.16	0.14	0.47	0.31	0.23	−0.18	0.39	0.22	0.38	0.23	0.20	0.55	0.21
Culm length (cm)	0.63	0.20	−0.13	0.47	0.61	0.42	0.20	0.56	0.63	0.63	−0.03	−0.31	0.31	0.31
Culm node number	0.70	0.00	0.38	0.53	0.73	0.36	0.34	0.52	0.56	0.58	0.10	0.31	0.49	0.34
Internode length (cm)	0.51	0.40	0.36	0.53	0.66	0.37	0.33	0.70	0.59	0.40	−0.17	0.24	0.33	0.15
Culm dry weight (g)	0.78	0.31	0.32	0.24	0.67	0.51	0.35	0.40	0.64					
Culm volume (cm^3)	0.72	0.34	−0.01	0.10	0.66	0.49	0.16	0.38	0.59	0.64	−0.16	0.04	0.59	0.06
Culm density (g/cm^3)	0.56	0.21	0.26		0.74	0.06	0.25	0.63	0.42					
Diameter of basal internode (mm)	0.77	0.29	0.16	0.16	0.66	0.51	0.40	0.36	0.64	0.54	−0.12	0.05	0.47	0.26
Diameter of topmost internode (mm)	0.57	0.42	−0.16	−0.15	0.57	0.42	−0.03	0.30	0.59	0.52	−0.15	0.16	0.64	0.13
Total number of culms	0.37	0.20	−0.06	−0.17	0.38	0.35	0.42	−0.15	0.55	0.12	0.33	−0.04	0.06	0.30

(Continues)

TABLE 4 (Continued)

	HU				NEF				UI	KNU		ZJU		
	S Japan	N Japan	Korea/N China	Yangtze-Qinling	S Japan	N Japan	Korea/N China	Yangtze-Qinling	S Japan	S Japan	Korea/N China	Korea/N China	Yangtze-Qinling	SE China/tropical
Proportion of reproductive culms	0.29	0.07	0.31		0.50	0.30	0.12	0.56	0.64			0.45	0.32	0.23
Culms per footprint (#/cm^2)	0.76	0.54	0.07	−0.17	0.48	0.19	−0.19	0.15	0.64	0.46	−0.10	0.04	−0.02	0.53

Abbreviations: SNP, single nucleotide polymorphism; BLUP, best linear unbiased predictor; HU, Hokkaido University; NEF, New Energy Farms; UI, University of Illinois at Urbana-Champaign; KNU, Kangwon National University; ZJU, Zhejiang University; $N_{genotypes}$, number of genotypes with phenotypic data for a given genetic group and field trial location.

Similarly, genomic selection analyses of multilocation BLUPs resulted in moderate to high prediction accuracies for year 3 biomass yield, ranging from 0.17 when all sites were included and population structure was controlled for (Table 3) to 0.57 for KNU + ZJU when we did not control for population structure (Table S2). Using data from temperate sites only and controlling for population structure, prediction accuracy for year 3 yield was 0.35 using BLUPs for genotype-within-genetic-group from Equation 4 and 0.31 using residuals of genotype BLUPs fitted to the genetic group from Equation 5 (Table 3; Figure 2), indicating that genomic selection across similar environments on individual genetic groups will be effective. Moreover, these prediction accuracies increased to 0.50 and 0.49, respectively, when a selection index for yield was used in place of direct selection for yield (Table 3), demonstrating the benefit of measuring yield component traits and including them in prediction models. In practice, genomic selection is likely to focus on breeding within genetic groups, and thus the prediction accuracies using genotype-within-genetic-group BLUPs and residuals of genotype BLUPs fitted on genetic groups provide useful estimates of the efficacy of genomic selection in real-world breeding. When genomic prediction was performed within genetic groups, prediction accuracy was variable due to small sample size (Tables 4 and 5). Nevertheless, we found that S Japan, the most consistently high-yielding genetic group across field trial locations, also had particularly high genomic prediction accuracies (Tables 4 and 5), indicating strong potential for genomic selection within this group. If the moderate to high genomic prediction accuracies from this study are confirmed in subsequent progeny populations, then we will be highly confident in the efficacy of genomic selection in *M. sinensis* and its advantage relative to phenotypic selection for efficiently breeding improved biomass cultivars.

In some cases, low or negative genomic prediction accuracies were observed for residuals after genotype BLUPs were fitted to genetic groups (Equation 5) or for genotype-within-genetic group BLUPs (Equations 3 and 4), despite moderate or high prediction accuracies using genotype BLUPs (Table 3), indicating that in those cases most of the prediction accuracy using genotype BLUPs was dependent on phenotypic differences between genetic groups, rather than variance within groups. For example, at ZJU, some phenotypic differences among genetic groups were especially pronounced, with the Sichuan and SE China groups being high yielding and having long, thick culms, whereas surviving genotypes in the ornamental group were shorter and had many more culms than other groups (Clark et al., 2019). For basal circumference, outliers were present in the NEF and ZJU datasets after Box–Cox transformation and genotype-within-genetic-group

TABLE 5 Genomic prediction accuracies within genetic groups for traits measured in *Miscanthus sinensis* in year 3 at five northern field trial locations (HU, NEF, CSU, UI, and KNU; 37.9–43.1°N) and one southern location (ZJU; 29.8°N). Phenotypic values were Box–Cox transformed, then genotypic BLUPs were calculated across locations (Equation 2) and used as the response variable in genomic prediction. The analyses were based on 46,177 SNP markers and 568 accessions. Genetic groups were excluded from analysis when fewer than 50 genotypes had phenotypic data

	HU + NEF + CSU + UI + KNU				HU + NEF +CSU + UI + KNU + ZJU				
	S Japan	N Japan	Korea/N China	Yangtze-Qinling	S Japan	N Japan	Korea/N China	Yangtze-Qinling	SE China/tropical
$N_{genotypes}$	141	84	156	71	142	84	156	74	77
Selection index (Equation 7)	0.73	0.44	0.42	0.43	0.79	0.57	0.43	0.59	−0.11
Dry biomass yield (g/plant)	0.56	0.28	0.39	0.33	0.58	0.28	0.23	0.48	0.21
Compressed circumference (cm)	0.56	0.19	0.29	0.35	0.56	0.32	0.30	0.47	0.18
Basal circumference (cm)	0.59	0.32	0.21	0.16	0.60	0.28	0.06	0.30	0.01
Compressed circumference/basal circumference	0.42	0.17	−0.02	0.40	0.42	0.26	0.10	0.50	−0.08
Culm length (cm)	0.69	0.38	0.22	0.49	0.70	0.35	0.27	0.38	0.35
Culm node number	0.77	0.36	0.38	0.57	0.78	0.42	0.37	0.59	0.16
Internode length (cm)	0.71	0.48	0.35	0.70	0.69	0.42	0.34	0.69	0.27
Culm volume (cm^3)	0.75	0.48	0.28	0.23	0.76	0.46	0.34	0.27	−0.20
Diameter of basal internode (mm)	0.76	0.47	0.35	0.30	0.77	0.52	0.37	0.41	−0.25
Diameter of topmost internode (mm)	0.65	0.46	−0.03	0.16	0.67	0.48	0.09	0.17	−0.27
Total number of culms	0.50	0.40	0.35	−0.17	0.51	0.25	0.34	−0.15	0.38
Culms per footprint (#/cm^2)	0.66	0.35	0.07	0.12	0.66	0.35	0.05	0.08	0.12

Abbreviations: SNP, single nucleotide polymorphism; BLUP, best linear unbiased predictor; HU, Hokkaido University; NEF, New Energy Farms; CSU, Colorado State University; UI, University of Illinois at Urbana-Champaign; KNU, Kangwon National University; ZJU, Zhejiang University; $N_{genotypes}$, number of genotypes with phenotypic data for a given genetic group and combination of field trial locations.

BLUP calculation, and inaccurate GEBVs for these outliers accounted for the low prediction accuracy at those sites. Both cases suggest that inspection of the relationship between phenotypic values and GEBVs, beyond a single correlation statistic, is important for predicting the efficacy of genomic selection.

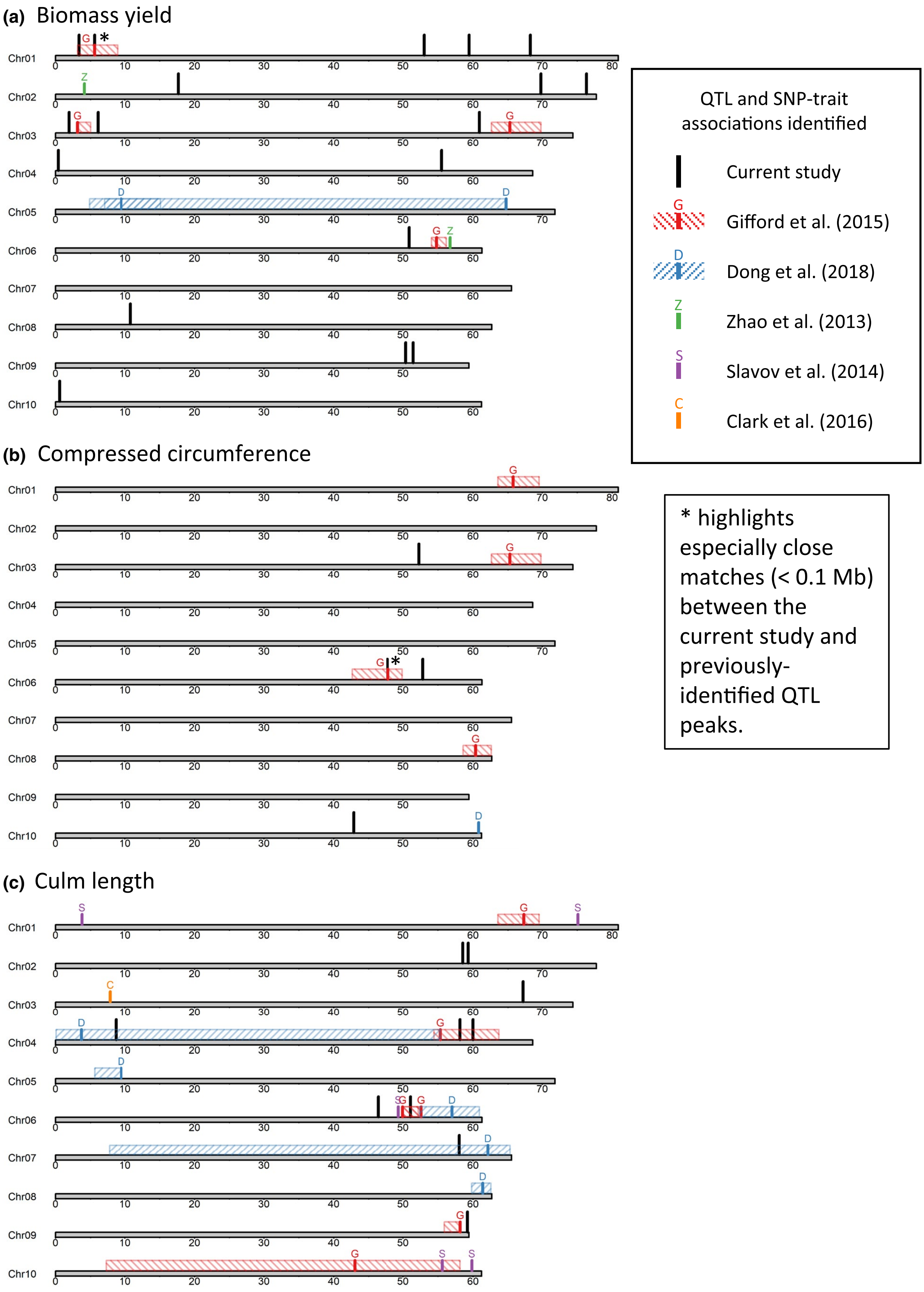
Several methods have been used by others to control for or assess the effect of population structure on genomic prediction. Guo et al. (2014) decomposed the relationship matrix into eigenvectors before using it for prediction, then estimated variance attributable to the first *n* eigenvectors, which was assumed to represent population structure. Azevedo et al. (2017) used the first several eigenvectors of the relationship matrix as covariates in the prediction model. Fiedler et al. (2018) assigned entire genetic groups to the training or validation set. Arruda et al. (2015) used *k*-means clustering to identify closely related individuals and to make sure they were not in both the training and validation sets. Lipka et al. (2014) fitted phenotypic values to the first several principal components of the marker data and used the residuals as the response variable in genomic prediction. Because we have previously identified genetic groups within *M. sinensis* (Clark et al., 2014, 2015) and evaluated phenotypic differences among these groups in this set of field trials (Clark et al., 2019), we corrected for population structure by removing variance attributable to genetic groups by (a) performing genomic prediction within individual groups (Tables 4 and 5), (b) estimating BLUPs for genotype-within-genetic-group from Equations 3 and 4 (*G*(*D*); Table 3), or (c) analyzing residuals of genotype BLUPs fitted to genetic group (*R*; Equation 5). Prediction accuracies were typically lower for the methods that accounted for population structure relative to analyses that did not account for structure, which was consistent with expectations because failure to account for population structure has been shown to bias estimates upwards (Fiedler et al., 2018; Guo et al., 2014; Riedelsheimer et al., 2012). For biomass yield, genomic prediction accuracies were mostly moderate for each of the three methods we used to control for population structure (Tables 3–5). Including multiple locations in the analyses, and thus having greater phenotypic sampling, was advantageous. The genotype-within-genetic-group method and the residuals method typically gave similar results to each other but the former often had slightly higher prediction accuracies than the latter (Table 3). These small differences may be due to Equation 4 controlling for genetic-group-by-location effects, whereas Equation 5 did not. It will be interesting to see if tests of future generations of *Miscanthus* confirm the small difference in prediction accuracies observed for these two methods in this study.

Our high genomic prediction accuracies and number of detected SNP–trait associations are notable given the relatively modest number of SNPs evaluated and the low linkage disequilibrium expected for *M. sinensis*, due to its self-incompatibility, undomesticated status, wind dispersal of pollen and seed, and large population size. Indeed, Slavov et al. (2014) found that linkage disequilibrium in *M. sinensis* decayed after several hundred base pairs for most SNP pairs (although they note that their ability to estimate linkage disequilibrium was limited by the low proportion of the genome covered by RAD tags, similar to our study) and yet they too were also able to identify significant associations for many traits. Given this low level of linkage disequilibrium, we might have expected that most blocks of linkage disequilibrium in the *M. sinensis* genome would not contain any SNPs from our set of 46,177 if the SNPs had been dispersed completely at random. However, our SNP markers were highly concentrated in the protein-coding portions of the genome. Out of 34,211 genes annotated in the *S. bicolor* genome v. 3.1, 14,062 of them were within 1 kb of a SNP in our dataset (20,611 SNPs near at least one gene; 11,013 genes actually contained a SNP from our dataset and 16,825 SNPs were within genes; see expanded version of Data S2 available at https://doi.org/10.13012/B2IDB-0790815_V3). Given the whole genome duplication of *Miscanthus* with respect to its common ancestor with sorghum, we estimate that at least one fifth of *M. sinensis* genes (~14K out of ~68K) were in linkage disequilibrium with one or more SNPs in our dataset. Moreover, hundreds of significant SNP–trait associations were identified and moderate to high genomic prediction accuracies were obtained with the current dataset. However, as only a minority of all *M. sinensis* genes are likely in linkage disequilibrium with the current marker set, substantially increasing the numbers of SNPs would be expected to greatly increase the effectiveness of the GWA and genomic prediction analyses.

We found that when the same SNP was significantly associated with multiple traits, those traits tended to have strong genetic correlation, suggesting that in some cases one trait can serve as a proxy for another in GWA. For example, UIMiscanthus105065 was significantly associated with yield and compressed circumference, the best predictor of yield at ZJU in year 2, as well as basal circumference at ZJU and KNU in year 3, where the genetic correlation with yield was 0.76 and 0.61, respectively (Clark et al., 2019). Thirteen SNPs had associations with multiple traits relating to culm dimensions and number of culms, including diameter of basal internode, diameter of topmost internode, culm length, culm node number, internode length, culm dry weight, culm volume, total number of culms, and culms per footprint. Diameter of basal internode, diameter of topmost internode, culm dry weight, and culm volume nearly always had moderate to strong positive genetic correlations with each other (>0.6) and moderate negative genetic correlations with culms per footprint (Clark et al., 2019), which made sense given that fewer thick culms can fit into a given area than thin culms. Of the 13 SNPs, 6 had significant associations at multiple nonoverlapping field sites and/or site combinations, consistent with the high multilocation heritabilities of these traits (UIMiscanthus020125, 022671, 092590, 097427, 105978, and 117199; Clark et al., 2019). Thicker, larger stems were typically associated with higher yielding plants (Clark et al., 2019), and thus, we expect these SNPs to be useful for selection and breeding. Lastly, we identified a SNP associated with the ratio of compressed to basal circumference and culms per footprint at ZJU in year 3 (UIMiscanthus098471), where both of these traits relate to how much of the area occupied by the plant is filled in with culms rather than empty space. Overall, the identification of SNPs significantly associated with multiple yield-component traits suggests that the associations are biologically meaningful and may be used for identification of candidate genes and for breeding.

Throughout the *Miscanthus* genome, 80 quantitative trait loci (QTL) identified by five previous studies (Clark et al., 2016; Dong et al., 2018; Gifford, Chae, Swaminathan, Moose, & Juvik, 2015; Slavov et al., 2014; Zhao et al., 2013) included or were near to one or more significant

FIGURE 3 Locations of significantly associated single nucleotide polymorphisms (SNPs) and quantitative trait locus (QTL) from this study and others for biomass yield, compressed circumference, and culm length in *Miscanthus* with respect to the *Sorghum bicolor* reference genome. QTL peaks and 95% confidence intervals are indicated for Gifford et al. (2015) and Dong et al. (2018), and locations of single associated markers are indicated for all other studies


(a) Biomass yield
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
G
Z
D
S
C
*
0
10
20
30
40
50
60
70
80
(b) Compressed circumference
(c) Culm length
QTL and SNP-trait associations identified
Current study
Gifford et al. (2015)
Dong et al. (2018)
Zhao et al. (2013)
Slavov et al. (2014)
Clark et al. (2016)
* highlights especially close matches (< 0.1 Mb) between the current study and previously-identified QTL peaks.

SNPs from the current study (Data S2). Out of the 36 QTL identified by Gifford et al. (2015) for traits that were also measured in the current study, 10 contained a SNP from our study significantly associated with the same trait and 33 contained at least one SNP from our study associated with any trait. Out of the 14 QTL identified by Gifford et al. (2015) for compressed circumference, basal circumference, or ratio of compressed to basal circumference, 9 contained SNPs from our study significantly associated with the total number of culms, which had moderate genetic correlations with these traits at most sites (Clark et al., 2019) and logically should influence them. UIMiscanthus093867, which was significantly associated with yield in our study, was only 3.4 kb from the peak of Gifford et al.'s (2015) yield QTL Y4 and number of culms QTL NT4 on sorghum chromosome 1 (Figure 3a; Data S2). Similarly, UIMiscanthus030627, significantly associated with compressed circumference in our study, was 77.5 kb from the peak of compressed circumference QTL CC2 identified by Gifford et al. (2015) on sorghum chromosome 6 (Figure 3b; Data S2). Out of the 47 joint-population meta-QTL identified by Dong et al. (2018) for traits included in this study, 7 spanned regions that included a SNP from our study for the same trait, and 34 included a SNP from our study significant for any trait. Additionally, out of nine QTL identified by Dong et al. (2018) for diameter of topmost internode, two spanned SNPs from our study for diameter of basal internode (strong positive genetic correlation; Clark et al., 2019), four spanned SNPs from our study for total number of culms (weak to moderate negative genetic correlation), and four spanned SNPs from our study for number of culms per area (moderate to strong negative genetic correlation). Concurrence between QTL identified in our study and others suggests confidence in their identification and indicates regions of the genome to prioritize further for identifying candidate genes.

Significant SNPs from our study that were near significant SNPs or QTL peaks from other studies were often associated with multiple yield-related traits. On *S. bicolor* chromosome 10, Slavov et al. (2014) identified a SNP associated with height of the tallest stem at 59.9 Mb (Sb10g029835), which is close to our SNP UIMiscanthus018832 (60.1 Mb) that was significantly associated with both culm node number and culm dry weight at NEF in year 3 (Table S1). Among the significant SNPs from our study that were closest to the QTL peaks identified by Gifford et al. (2015), UIMiscanthus117119, 030627, and 007880 were all associated with multiple traits (Data S2; Table S1).

4.2 | Candidate genes identified

Many of the significant SNP–trait associations identified in this study, especially those within previously known QTL, were near or in genes with a previously determined function in *Arabidopsis* and/or rice that suggested they may be the causative gene for the traits we observed in *M. sinensis* (Table S1; Data S2). For example, UIMiscanthus020125, which was associated with culm diameter and total number of culms in our study, as well as being within a previously identified QTL on *S. bicolor* chromosome 6 for number of culms (Gifford et al., 2015), codes for a synonymous mutation in the second exon of Sobic.006G034100, a methyl-CpG-binding protein expressed in juvenile and mature stems, lower leaves at grain maturity, and panicles, according to GeneAtlas data available at Phytozome (DOE-JGI, https://phytozome.jgi.doe.gov). Given that methyl-CpG-binding proteins have been observed to interpret DNA methylation in plants (Zemach & Grafi, 2007), regulating lateral branching in at least one case (Peng, Cui, Bi, & Rothstein, 2006), we hypothesize that this gene may be involved in epigenetic regulation of the observed trade-off in *M. sinensis* plants between the production of many thin stems or fewer, thicker stems. Similarly, on chromosome 3, UIMiscanthus112990, which was associated with total number of culms in the current study and within a QTL region identified by Gifford et al. (2015) for the same trait, is within the intron of the WOX transcription factor Sobic.003G336600 that is expressed in the panicle, stem, and leaf. WOX transcription factors have been found to regulate cell division and differentiation, including lateral organ formation in maize, *Arabidopsis*, and petunia (Van Der Graaff, Laux, & Rensing, 2009), which is consistent with the observed association for number of culms in *M. sinensis*. On chromosome 5, UIMiscanthus019070, which is 800 bp downstream of Sobic.005G132000, a transcription factor that regulates the expression of genes that respond to auxins (Guilfoyle & Hagen, 2007), was associated with the number of culms at HU in year 3, and is within overlapping QTL for basal circumference identified by Dong et al. (2018) and Gifford et al. (2015). Given the important role of auxin in tillering (Hussien et al., 2014), this is another promising candidate gene. For each significant SNP, Data S2 lists any sorghum genes containing the SNP as well as any sorghum genes within 1 kb of the SNP, along with any known gene functions from *Arabidopsis* or *Oryza*, and can be used as a resource for mining genes for further study.

5 | CONCLUSIONS

Genomic prediction has the potential to accelerate the breeding of *M. sinensis* several-fold in the immediate future. Large populations of seedlings can be screened with RAD-seq and subjected to genomic selection. Given the genomic prediction accuracies that we obtained, we recommend using genomic selection to identify the top percentage of genotypes for predicted yield breeding values, which can then be used for rapid cycling of generations to improve genetic gain per year relative to phenotypic selection. We expect that rapid and substantial genetic gains for biomass yield in *M. sinensis* will be obtained by implementing these new methods.

ACKNOWLEDGEMENTS

This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), Grant Nos. DE-SC0006634 and DE-SC0018420. New Energy Farms provided in-kind support. We especially thank our field crews at all institutions for many hours planting and maintaining field trials and taking measurements.

ORCID

Lindsay V. Clark https://orcid.org/0000-0002-3881-9252
Alexander E. Lipka https://orcid.org/0000-0003-1571-8528
Toshihiko Yamada https://orcid.org/0000-0002-7845-6556
Stephen P. Long https://orcid.org/0000-0002-8501-7164

REFERENCES

Arnoult, S., & Brancourt-Hulmel, M. (2015). A review on *Miscanthus* biomass production and composition for bioenergy use: Genotypic and environmental variability and implications for breeding. *BioEnergy Research*, *8*(2), 502–526. https://doi.org/10.1007/s12155-014-9524-7

Arruda, M. P., Brown, P. J., Lipka, A. E., Krill, A. M., Thurber, C., & Kolb, F. L. (2015). Genomic selection for predicting head blight resistance in a wheat breeding program. *The Plant Genome*, *8*(3). https://doi.org/10.3835/plantgenome2015.01.0003

Azevedo, C. F., Resende, M. D. V. D., Silva, F. F. E., Nascimento, M., Viana, J. M. S., & Valente, M. S. F. (2017). Population structure correction for genomic selection through eigenvector covariates. *Crop Breeding and Applied Biotechnology*, *17*(4), 350–358. https://doi.org/10.1590/1984-70332017v17n4a53

Baker, R. J. (1986). *Selection indices in plant breeding*. Boca Raton, FL: CRC Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. http://www.jstor.org/stable/2346101.

Bian, Y., & Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity*, *118*(6), 585–593. https://doi.org/10.1038/hdy.2017.4

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. http://www.jstor.org/stable/2984418

Clark, L. V., Brummer, J. E., Głowacka, K., Hall, M. C., Heo, K., Peng, J., … Sacks, E. J. (2014). A footprint of past climate change on the diversity and population structure of *Miscanthus sinensis*. *Annals of Botany*, *114*(1), 97–107. https://doi.org/10.1093/aob/mcu084

Clark, L. V., Dwiyanti, M. S., Anzoua, K. G., Brummer, J. E., Ghimire, B. K., Głowacka, K., … Sacks, E. J. (2019). Biomass yield in a genetically diverse *Miscanthus sinensis* germplasm panel evaluated at five locations revealed individuals with exceptional potential. *GCB Bioenergy*, https://doi.org/10.1111/gcbb.12606

Clark, L. V., Dzyubenko, E., Dzyubenko, N., Bagmet, L., Sabitov, A., Chebukin, P., … Sacks, E. J. (2016). Ecological characteristics and in situ genetic associations for yield-component traits of wild *Miscanthus* from eastern Russia. *Annals of Botany*, *118*(5), 941–955. https://doi.org/10.1093/aob/mcw137

Clark, L. V., Stewart, J. R., Nishiwaki, A., Toma, Y., Kjeldsen, J. B., Jørgensen, U., … Sacks, E. J. (2015). Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. *Journal of Experimental Botany*, *66*(14), 4213–4225. https://doi.org/10.1093/jxb/eru511

Clifton-Brown, J. C., Chiang, Y.-C., & Hodkinson, T. R. (2008). *Miscanthus*: Genetic resources and breeding potential to enhance bioenergy production. In W. Vermerris (Ed.), *Genetic improvement of bioenergy crops* (pp. 273–294). New York, NY: Springer. https://doi.org/10.1007/978-0-387-70805

Clifton-Brown, J., Harfouche, A., Casler, M. D., Dylan Jones, H., Macalpine, W. J., Murphy-Bokern, D., … Lewandowski, I. (2019). Breeding progress and preparedness for mass-scale deployment of perennial lignocellulosic biomass crops switchgrass, *Miscanthus*, willow, and poplar. *GCB Bioenergy*, *11*(1), 118–151. https://doi.org/10.1111/gcbb.12566

Clifton-Brown, J. C., & Lewandowski, I. (2000). Overwintering problems of newly established *Miscanthus* plantations can be overcome by identifying genotypes with improved rhizome cold tolerance. *New Phytologist*, *148*(2), 287–294. https://doi.org/10.1046/j.1469-8137.2000.00764.x

Clifton-Brown, J. C., Lewandowski, I., Andersson, B., Basch, G., Christian, D. G., Kjeldsen, J. B., … Teixeira, F. (2001). Performance of 15 *Miscanthus* genotypes at five sites in Europe. *Agronomy Journal*, *93*(5), 1013–1019. https://doi.org/10.2134/agronj2001.9351013x

Crossa, J., Burgueño, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S. A., Lillemo, M., … Ortiz, R. (2007). Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics*, *177*(3), 1889–1913. https://doi.org/10.1534/genetics.107.078659

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., … Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, *112*(1), 48–60. https://doi.org/10.1038/hdy.2013.16

Davey, C. L., Robson, P., Hawkins, S., Farrar, K., Clifton-Brown, J. C., Donnison, I. S., & Slavov, G. T. (2017). Genetic relationships between spring emergence, canopy phenology and biomass yield increase the accuracy of genomic prediction in *Miscanthus*. *Journal of Experimental Botany*, *68*(18), 5093–5102. https://doi.org/10.1093/jxb/erx339

Dong, H., Liu, S., Clark, L. V., Sharma, S., Gifford, J. M., Juvik, J. A., … Sacks, E. J. (2018). Genetic mapping of biomass yield in three interconnected *Miscanthus* populations. *GCB Bioenergy*, *10*(3), 165–185. https://doi.org/10.1111/gcbb.12472

Dong, H., Green, S. V., Nishiwaki, A., Yamada, T., Stewart, J. R., Deuter, M., & Sacks, E. J. (2019). Winter hardiness of *Miscanthus* (I): Overwintering ability and yield of new *Miscanthus* × *giganteus* genotypes in Illinois and Arkansas. *GCB Bioenergy*, *11*(5), 691–705. https://doi.org/10.1111/gcbb.12588

Dwiyanti, M. S., Stewart, J. R., & Yamada, T. (2013). Germplasm resources of *Miscanthus* and their application in breeding. In M. C.

Saha, H. S. Bhandari, & J. H. Bouton (Eds.), *Bioenergy feedstocks: Breeding and genetics* (pp. 49–66). Oxford, UK: John Wiley & Sons, Inc.. https://doi.org/10.1002/9781118609477.ch4

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome Journal*, *4*(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587. Retrieved from http://www.genetics.org/content/164/4/1567

Fiedler, J. D., Lanzatella, C., Edmé, S. J., Palmer, N. A., Sarath, G., Mitchell, R., & Tobias, C. M. (2018). Genomic prediction accuracy for switchgrass traits related to bioenergy within differentiated populations. *BMC Plant Biology*, *18*(1), 1–16. https://doi.org/10.1186/s12870-018-1360-z

Gifford, J. M., Chae, W. B., Swaminathan, K., Moose, S. P., & Juvik, J. A. (2015). Mapping the genome of *Miscanthus sinensis* for QTL associated with biomass productivity. *GCB Bioenergy*, *7*(4), 797–810. https://doi.org/10.1111/gcbb.12201

Głowacka, K. (2011). A review of the genetic study of the energy crop *Miscanthus*. *Biomass and Bioenergy*, *35*(7), 2445–2454. https://doi.org/10.1016/j.biombioe.2011.01.041

Głowacka, K., Clark, L. V., Adhikari, S., Peng, J., Stewart, J. R., Nishiwaki, A., … Sacks, E. J. (2015). Genetic variation in *Miscanthus* × *giganteus* and the importance of estimating genetic distance thresholds for differentiating clones. *GCB Bioenergy*, *7*(2), 386–404. https://doi.org/10.1111/gcbb.12166

Guilfoyle, T. J., & Hagen, G. (2007). Auxin response factors. *Current Opinion in Plant Biology*, *10*(5), 453–460. https://doi.org/10.1016/j.pbi.2007.08.014

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., … Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. *Theoretical and Applied Genetics*, *127*(3), 749–762. https://doi.org/10.1007/s00122-013-2255-x

Heaton, E. A., Clifton-Brown, J., Voigt, T. B., Jones, M. B., & Long, S. P. (2004). *Miscanthus* for renewable energy generation: European Union experience and projections for Illinois. *Mitigation and Adaptation Strategies for Global Change*, *9*(4), 433–451. https://doi.org/10.1023/B:MITI.0000038848.94134.be

Hussien, A., Tavakol, E., Horner, D. S., Muñoz-Amatriaín, M., Muehlbauer, G. J., & Rossini, L. (2014). Genetics of tillering in rice and barley. *The Plant Genome*, *7*. https://doi.org/10.3835/plantgenome2013.10.0032

International Rice Research Institute. (2002). Standard evaluation system for rice. Retrieved from http://www.knowledgebank.irri.org/images/docs/rice-standard-evaluation-system.pdf

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jost, L. (2008). G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, *17*(18), 4015–4026. https://doi.org/10.1111/j.1365-294X.2008.03887.x

Kaiser, C. M., Clark, L. V., Juvik, J. A., Voigt, T. B., & Sacks, E. J. (2015). Characterizing a *Miscanthus* germplasm collection for yield, yield components, and genotype × environment interactions. *Crop Science*, *55*(5), 1978–1994. https://doi.org/10.2135/cropsci2014.12.0808

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lipka, A. E., Lu, F., Cherney, J. H., Buckler, E. S., Casler, M. D., & Costich, D. E. (2014). Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. *PLoS ONE*, *9*(11), e112227. https://doi.org/10.1371/journal.pone.0112227

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., … Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Liu, H.-J., & Yan, J. (2019). Crop genome-wide association study: A harvest of biological relevance. *The Plant Journal*, *97*(1), 8–18. https://doi.org/10.1111/tpj.14139

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., … Costich, D. E. (2013). Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genetics*, *9*(1), e1003215. https://doi.org/10.1371/journal.pgen.1003215

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. Retrieved from http://www.genetics.org/content/157/4/1819

Nie, G., Huang, L., Zhang, X., Taylor, M., Jiang, Y., Yu, X., … Zhang, Y. (2016). Marker-trait association for biomass yield of potential biofuel feedstock *Miscanthus sinensis* from southwest China. *Frontiers in Plant Science*, *7*, 802. https://doi.org/10.3389/fpls.2016.00802

Owens, B. F., Lipka, A. E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., … Rocheford, T. (2014). A foundation for provitamin a biofortification of maize: Genome-wide association and genomic prediction models of carotenoid levels. *Genetics*, *198*(4), 1699–1716. https://doi.org/10.1534/genetics.114.169979

Peng, M., Cui, Y., Bi, Y. M., & Rothstein, S. J. (2006). AtMBD9: A protein with a methyl-CpG-binding domain regulates flowering time and shoot branching in *Arabidopsis*. *Plant Journal*, *46*(2), 282–296. https://doi.org/10.1111/j.1365-313X.2006.02691.x

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., … Jannink, J.-L. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome Journal*, *5*(3), 103. https://doi.org/10.3835/plantgenome2012.06.0006

Resende, M. F. R., Munoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., … Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics*, *190*(4), 1503–1510. https://doi.org/10.1534/genetics.111.137026

Rice, B., & Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *The Plant Genome*, *12*(1), 180052. https://doi.org/10.3835/plantgenome2018.07.0052

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., … Melchinger, A. E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics*, *44*(2), 217–220. https://doi.org/10.1038/ng.1033

Sacks, E. J., Juvik, J. A., Lin, Q., Stewart, J. R., & Yamada, T. (2013). The gene pool of *Miscanthus* species and its improvement. In A. H. Paterson (Ed.), *Genomics of the Saccharinae* (pp. 73–101). New York, NY: Springer New York: https://doi.org/10.1007/978-1-4419-5947-8

Scheben, A., & Edwards, D. (2018). Bottlenecks for genome-edited crops on the road from lab to farm. *Genome Biology*, *19*(1), 5–11. https://doi.org/10.1186/s13059-018-1555-5

Slavov, G. T., Davey, C. L., Bosch, M., Robson, P. R. H., Donnison, I. S., & Mackay, I. J. (2019). Genomic index selection provides a pragmatic framework for setting and refining multi-objective breeding targets in *Miscanthus*. *Annals of Botany*, in Press., https://doi.org/10.1093/aob/mcy187

Slavov, G. T., Nipper, R., Robson, P., Farrar, K., Allison, G. G., Bosch, M., … Jensen, E. (2014). Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*. *New Phytologist*, *201*(4), 1227–1239. https://doi.org/10.1111/nph.12621

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4), 395–408. https://doi.org/10.1038/hdy.2015.113

Van Der Graaff, E., Laux, T., & Rensing, S. A. (2009). The WUS homeobox-containing (WOX) protein family. *Genome Biology*, *10*, 248. https://doi.org/10.1186/gb-2009-10-12-248

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wei, X., Jackson, P. A., Hermann, S., Kilian, A., Heller-Uszynska, K., & Deomano, E. (2010). Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker–trait associations in sugarcane. *Genome*, *53*(11), 973–981. https://doi.org/10.1139/G10-050

Yan, J., Chen, W., Luo, F., Ma, H., Meng, A., Li, X., … Sang, T. (2012). Variability and adaptability of *Miscanthus* species evaluated for energy crop domestication. *GCB Bioenergy*, *4*(1), 49–60. https://doi.org/10.1111/j.1757-1707.2011.01108.x

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., … Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Zemach, A., & Grafi, G. (2007). Methyl-CpG-binding domain proteins in plants: Interpreters of DNA methylation. *Trends in Plant Science*, *12*(2), 80–85. https://doi.org/10.1016/j.tplants.2006.12.004

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., … Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/ng.546

Zhao, H., Wang, B., He, J., Yang, J., Pan, L., Sun, D., & Peng, J. (2013). Genetic diversity and population structure of *Miscanthus sinensis* germplasm in China. *PLoS ONE*, *8*(10), e75672. https://doi.org/10.1371/journal.pone.0075672

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Clark LV, Dwiyanti MS, Anzoua KG, et al. Genome-wide association and genomic prediction for biomass yield in a genetically diverse *Miscanthus sinensis* germplasm panel phenotyped at five locations in Asia and North America. *GCB Bioenergy*. 2019;11:988–1007. https://doi.org/10.1111/gcbb.12620