Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass Miscanthus sinensis

Gancho T. Slavov1, Rick Nipper2, Paul Robson1, Kerrie Farrar1, Gordon G. Allison1, Maurice Bosch1, John C. Clifton-Brown1, Iain S. Donnison1 and Elaine Jensen1

1Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3EB, UK; 2Floragenex Inc., Eugene, OR 97403, USA

Summary

- Increasing demands for food and energy require a step change in the effectiveness, speed and flexibility of crop breeding. Therefore, the aim of this study was to assess the potential of genome-wide association studies (GWASs) and genomic selection (i.e. phenotype prediction from a genome-wide set of markers) to guide fundamental plant science and to accelerate breeding in the energy grass Miscanthus.
- We generated over 100 000 single-nucleotide variants (SNVs) by sequencing restriction site-associated DNA (RAD) tags in 138 Miscanthus sinensis genotypes, and related SNVs to phenotypic data for 17 traits measured in a field trial.
- Confounding by population structure and relatedness was severe in naïve GWAS analyses, but mixed-linear models robustly controlled for these effects and allowed us to detect multiple associations that reached genome-wide significance. Genome-wide prediction accuracies tended to be moderate to high (average of 0.57), but varied dramatically across traits. As expected, predictive abilities increased linearly with the size of the mapping population, but reached a plateau when the number of markers used for prediction exceeded 10 000–20 000, and tended to decline, but remain significant, when cross-validations were performed across subpopulations.
- Our results suggest that the immediate implementation of genomic selection in Miscanthus breeding programs may be feasible.

Introduction

Cost-efficient genotyping protocols based on next-generation sequencing (Davey et al., 2011) have narrowed the gap between model and non-model plants, thereby creating great opportunities in crop breeding (Hamblin et al., 2011; Morrell et al., 2012). Two approaches that are likely to have particularly strong and widespread impacts are the dissection of complex traits through genome-wide association studies (GWASs; Cardon & Bell, 2001) and genome-wide phenotype prediction (genomic selection; Meuwissen et al., 2001). By informing fundamental science and applied breeding, respectively, these two approaches have the potential to bridge the molecular and statistical void between phenotype and genotype, thereby shedding light on key biological questions.

With increasing demands for food and energy, decreasing land base and changing environmental conditions (Foley et al., 2011; Valentine et al., 2012), the urgency to develop accelerated crop breeding strategies cannot be overstated. The use of lignocellulosic energy crops is one of many potential mitigating factors for this problem, but the widespread adoption of these crops has been slow and hampered by a number of technological and economic challenges, with major breakthroughs required both in terms of biomass yield on low-value lands and conversion efficiency (Sims et al., 2010; Feltus & Vandenbrink, 2012; Service, 2013). Genomic approaches (e.g. GWAS and genomic selection) can substantially facilitate these breakthroughs and are therefore among the focal points in this area of research (Rubin, 2008; Feltus & Vandenbrink, 2012; Slavov et al., 2013a).

The tropical C4 grass Miscanthus is a promising energy crop because of its broad adaptability, potentially very high yields and low requirements for agricultural inputs (Clifton-Brown et al., 2004, 2007; Hastings et al., 2009). However, Miscanthus species are undomesticated, and several breeding programs in East Asia, Europe and North America are targeting the accelerated development of hybrids and varieties that are high yielding, can be established and maintained at low cost and have cell wall characteristics that allow efficient conversion to fuels and products. In a recent study, we used a combination of phenotypic data from a replicated field trial and 120 molecular markers to delineate an experimental population of M. sinensis for GWAS and genome-wide prediction (Slavov et al., 2013b). This population had...
relatively weak substructure \((F_{ST} < 0.06)\), but captured high levels of genetic variation across a range of phenotypic traits related to phenology, biomass productivity and cell wall composition.

Building on this study, we: (1) sequenced restriction site-associated DNA (RAD) tags in 142 \(M.\) \(sinensis\) genotypes and generated over 100 000 ‘RAD-Seq’ single-nucleotide variants (SNVs); (2) confirmed patterns of putatively neutral population structure detected in our previous study, but also used the power of the RAD-Seq markers to add substantial new details, including the hypothetical geographical origins for a large number of accessions with unknown sampling locations; and (3) assessed the potential of GWASs and genome-wide prediction to guide biological discovery and accelerated breeding in \(Miscanthus\).

Materials and Methods

Plant materials, phenotyping and DNA extraction

We defined an experimental population of 145 \(M.\) \(sinensis\) Anders. genotypes based on previous analyses of single-nucleotide polymorphism (SNP) data (Slavov \textit{et al.}, 2013b). Of these, we attempted RAD sequencing for 142 and obtained robust data for 138 genotypes (see section on Sample clustering, population structure and relatedness), which were used for all subsequent analyses (Fig. 1). In 2005, our study population was planted at 1.5 × 1.5 m spacing in a replicated field trial located near Aberystwyth (Wales, UK), following a randomized complete block design, with one replicate per genotype in each of four blocks. The field trial has been described in greater detail in several previous studies (Allison \textit{et al.}, 2011; Jensen \textit{et al.}, 2011; Robson \textit{et al.}, 2012). Phenotyping and DNA extraction protocols have been described by Slavov \textit{et al.} (2013b). Briefly, 17 phenotypic traits reflecting (1) phenology, (2) morphology and biomass productivity, and (3) cell wall composition, were measured on plants in all four replicates of the trial 2–4 yr after establishment (Table 1, Fig. 2).

Phenotypic data analysis

Phenotypic data analyses have been described by Slavov \textit{et al.} (2013b). Briefly, mixed linear models were used to estimate variance components and best linear unbiased predictors (BLUPs) for each trait. We also calculated phenotypic (Pearson’s \(r\)) and genetic correlations among all pairs of traits (Fig. 2). Genetic correlations were calculated using two different approaches. For pairs of traits measured in the same year, we calculated genetic correlations from estimates of genetic covariance (i.e. using Eqn 9, as described by Howe \textit{et al.}, 2000). Alternatively, for pairs of traits measured in different years, we calculated genetic correlations from phenotypic correlations and estimates of broad sense heritability for each trait (i.e. Eqn 8, as described by Howe \textit{et al.}, 2000).

RAD sequencing

RAD library preparation was performed by Floragenex, Inc. (Eugene, OR, USA). Briefly, genomic DNA from each genotype was digested with the restriction endonuclease \(Pst\) I and processed into multiplexed RAD libraries following the methods described in previous studies (Baird \textit{et al.}, 2008; Stölting \textit{et al.}, 2013). \(Pst\) I adapters, each containing a unique 6-bp multiplex sequence index (barcode), were affixed to digested templates, polished and amplified via a polymerase chain reaction. The resulting RAD libraries were run on two Illumina (San Diego, CA, USA) HiSeq platforms at Ambyr Genetics (Aliso Viejo, CA, USA) and the Oregon State University Center for Genome Research and Biocomputing (Corvallis, OR, USA) using Illumina 1 × 100-bp chemistry.

Fig. 1 Geographical distribution and principal component (PC) analysis of population genetic structure for 138 \(Miscanthus\) \(sinensis\) genotypes using 14 073 single-nucleotide variant loci. The percentages of the total variation explained by each PC are shown in parentheses.
We used custom scripts and VCFtools (Danecek et al., 2011) to further filter SNV data based on alignment statistics, minor allele frequency (MAF) and conformity of genotype frequencies to Hardy–Weinberg expectations (Table 2). For initial data analyses aimed at validating data quality and detecting patterns of population genetic structure, we generated sets of SNVs that passed relatively stringent filtering criteria, but without restricting MAFs (i.e. ‘stringent’ filtering, Table 2). To enhance the power of GWAS analyses and to assess the accuracy of genome-wide prediction as a function of the number of markers used, we also generated larger sets of SNVs using ‘liberal’ filtering criteria (Table 2). We used the PLINK software tool (Purcell et al., 2007) to calculate the number of unlinked SNVs (\textit{plink} \ldots –indep-pairwise 50 5 0.2), as well as to estimate linkage disequilibrium (LD) (\textit{plink} \ldots –ld-window-r2).

Sample clustering, population structure and relatedness
We used the individual-based principal component analysis (PCA) approach of Patterson et al. (2006) to detect outliers and to characterize population genetic structure based on SNV data.
filtered using our ‘stringent’ criteria (Table 2), after eliminating one marker from each pair of loci linked at $r^2 \geq 0.8$ (i.e. based on genotypic correlation). Four of the 142 sequenced genotypes were identified as outliers along four different and highly significant axes of variation ($P < 10^{-5}$ from tests based on the Tracy–Widom distribution) based on the default settings of the smartpca program within the EIGENSOFT package (Patterson et al., 2006). These genotypes had not been identified as outliers or potential *M. sacchariflorus* hybrids in analyses based on microsatellite and SNP data (Slavov et al., 2013b), and we tentatively assumed that their detection as outliers based on RAD-Seq SNVs was an indication of inferior RAD-Seq data quality. These genotypes were therefore eliminated, and all subsequent analyses were performed using phenotypic and SNV data for the remaining 138 genotypes. However, by performing a subset of analyses based on all 142 genotypes, we also ensured that none of our major results depended on this decision.

Patterns of population structure detected using the PCA approach were also confirmed using the model-based clustering approach implemented in STRUCTURE (Pritchard et al., 2000; Falush et al., 2003, 2007), following the procedure used in a previous study (Slavov et al., 2013b). Briefly, we ran STRUCTURE using the default model parameters and varying the assumed number of genetic groups ($K$) from one to six. Each run consisted of 10 000 burn-in iterations and 10 000 data collection iterations. We used the DISTRACT program (Rosenberg, 2004) to visualize the results from 10 independent runs that had been aligned using the CLUMPP program (Jakobsson & Rosenberg, 2007). We also used the results from these runs to calculate the ad hoc statistic $\Delta K$, which tends to peak at the value of $K$ that corresponds to the highest hierarchical level of substructure (Evanno et al., 2005), using the online version of the STRUCTURE HARVESTER program (Earl & vonHoldt, 2012). Finally, we used the GCTA program (Yang et al., 2011) to calculate the genetic relationship matrix (Yang et al., 2010) among the 138 genotypes used in all analyses.

GWASs

We used the efficient mixed-model association expedited approach (EMMAX; Kang et al., 2010) to perform GWASs based on SNVs filtered using ‘liberal’ criteria (Table 2). To control for the confounding effects of cryptic relatedness and population structure, we incorporated the identity-by-state (IBS) kinship matrix (calculated on the basis of all markers using EMMAX) and the first two principal components of population structure (see Sample clustering, population structure and relatedness). This approach is widely used and is believed to provide adequate protection against both environmental and genetic confounding (Price et al., 2010; Vilhjálmsdóttir & Nordborg, 2013). We used data perturbation simulations (Yu et al., 2006) to estimate the statistical power of our EMMAX analyses, as well as to quantify the inflation of naïve estimates of the proportion of variance explained (PVE) by each detected association (Beavis, 1998; Allison et al., 2002; Ingvarsson et al., 2008). For each iteration of the simulations, we randomly chose an SNV from our
Statistics

No. of alleles
Min het reads
observed and expected heterozygosities.
filtering criteria based on alignments to an allele in a heterozygous genotype.
genotypes.
f and 
where 

\[ k \]

hlle

\[ f \]

ISf

IS|f

0.25

NA

Min het reads

MISSING (%), maximum percentage of missing genotype data allowed for a given locus.

Minor alleles, minimum number of copies of the minor allele among all genotypes.

Max F\text{IS}g, maximum deviation of observed genotype frequencies from Hardy–Weinberg expectations, \[ F_{IS} = 1 - H_o/H_e \], where \( H_o \) and \( H_e \) are the observed and expected heterozygosities.

Min het reads, minimum proportion of reads supporting the less frequent allele in a heterozygous genotype.

No. of alleles, number of SNV alleles detected.

No. of loci (Sorg), number of SNVs that passed all filtering criteria based on alignments to the Sorghum bicolor genome.

Ave MAF (Sorg), average minor allele frequency for SNVs that passed all filtering criteria based on alignments to the Sorghum bicolor genome.

Ave \( r^2 \text{IS} \) (Sorg), average linkage disequilibrium (\( r^2 \)), calculated as genotypic correlation) for pairs of loci with MAF \( \geq 0.10 \) that aligned within 1 kb of one another in the Sorghum bicolor genome.

No. of loci (Misc), number of SNVs that passed all filtering criteria based on alignments to an M. sinensis pseudo-reference.

Ave MAF (Misc), average minor allele frequency for SNVs that passed all filtering criteria based on alignments to an M. sinensis pseudo-reference. NA, not applicable.

Table 2 Filtering criteria for single-nucleotide variant (SNV) data from 138 Miscanthus sinensis genotypes

<table>
<thead>
<tr>
<th>Filtering criteria</th>
<th>Stringent</th>
<th>Liberal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q*</td>
<td>15</td>
<td>NA</td>
</tr>
<tr>
<td>Min depthb</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Min ave depthc</td>
<td>NA</td>
<td>6</td>
</tr>
<tr>
<td>Missing (%)d</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Minor allelesd</td>
<td>NA</td>
<td>3</td>
</tr>
<tr>
<td>Max F</td>
<td>isg</td>
<td>0.25</td>
</tr>
<tr>
<td>Min het readsf</td>
<td>0.05</td>
<td>NA</td>
</tr>
<tr>
<td>No. of allelesg</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

*Q, minimum Phred-like SNV quality score (Li et al., 2008).

*bMin depth, minimum number of reads.

*cMin ave depth, minimum average number of reads across all genotypes.

*dMissing (%), maximum percentage of missing genotype data allowed for a given locus.

*eMinor alleles, minimum number of copies of the minor allele among all genotypes.

*fMax F|is, maximum deviation of observed genotype frequencies from Hardy–Weinberg expectations, \( F_{IS} = 1 - H_o/H_e \), where \( H_o \) and \( H_e \) are the observed and expected heterozygosities.

*gMin het reads, minimum proportion of reads supporting the less frequent allele in a heterozygous genotype.

*hNo. of alleles, number of SNV alleles detected.

*iNo. of loci (Sorg), number of SNVs that passed all filtering criteria based on alignments to the Sorghum bicolor genome.

+jAve MAF (Sorg), average minor allele frequency for SNVs that passed all filtering criteria based on alignments to the Sorghum bicolor genome.

+kAve \( r^2 \text{IS} \) (Sorg), average linkage disequilibrium (\( r^2 \)), calculated as genotypic correlation) for pairs of loci with MAF \( \geq 0.10 \) that aligned within 1 kb of one another in the Sorghum bicolor genome.

+lNo. of loci (Misc), number of SNVs that passed all filtering criteria based on alignments to an M. sinensis pseudo-reference.

+mAve MAF (Misc), average minor allele frequency for SNVs that passed all filtering criteria based on alignments to an M. sinensis pseudo-reference. NA, not applicable.

To assess the relative severity of confounding across traits, we also ran models that only included the IBS kinship matrix, or did not include any terms to account for population structure and relatedness (i.e. simple linear regression of trait BLUP on individual SNVs). In addition, we also used the multi-locus mixed-model (MLMM) approach of Segura et al. (2012) to perform a forward–backward model selection procedure and to potentially improve the power, whilst also reducing the rate of false positives of our GWAS analyses. We allowed up to nine forward selection and backward elimination steps and performed the procedure twice, based on MLMMs with or without the first two principal components of population structure included as fixed effects.

**Genome-wide prediction**

We used the R package rrBLUP (Endelman, 2011) for genome-wide prediction using ridge regression. For a limited subset of analyses, we also used the Bayesian linear regression and the Bayesian LASSO approaches implemented in the BLR package (Perez et al., 2010), but the performance of genome-wide prediction using these approaches was consistently comparable or lower than that using ridge regression, even after including the kinship matrix in the Bayesian models. To assess the effects of varying the size of the training population, we used 2–10-fold random cross-validations (i.e. random allocations of genotypes to training and test populations), which were repeated 100 times for each set of parameters.

Consistent with recommendations on standardizing analytical procedures and benchmarking (Daetwyler et al., 2013), we quantified the performance of genome-wide prediction using three measures. First, we defined predictive ability (\( r \)) as Pearson’s correlation of BLUPs calculated directly from field data and those obtained from the marker data using ridge regression (Daetwyler et al., 2013). Second, we calculated prediction accuracies (\( \text{Accu} \)) by dividing predictive abilities by the square root of the broad-sense heritability (\( H \)) of the respective trait (Lagarra et al., 2008). Finally, we also recorded the intercepts (\( b_0 \)) and slopes (\( b_1 \)) of the simple linear regressions of BLUPs calculated from field data on those estimated using ridge regression. The last two measures, although not practically meaningful, can be indicative of model deficiencies and/or non-random partitioning of genotypes into training vs test populations (Daetwyler et al., 2013).

To assess the potential of improving the performance of genome-wide prediction by selecting the most informative markers, we compared predictive abilities between cross-validations in which 10, 100, 1000 or 10 000 markers were chosen randomly, based on their GWAS significance (i.e. selecting loci with the lowest GWAS \( P \) values) within the training population, or based on their rrBLUP-estimated effects within the training population.

---

\[
PVE = \frac{1}{1 + \frac{1}{2p(1-p)k^2}}, \quad \text{Eqn 1}
\]

where \( p \) is the estimated frequency of an arbitrarily chosen SNV allele and \( k \) is the simulated additive effect (\( a \)) divided by the standard deviation of the respective trait BLUPs (we used \( \text{AvgeSen.9} \)). We chose values of \( k \) so that the expected PVE would be between 0.01 and 0.20 in 0.01 increments, and performed 1000 simulations for each of these values. Naïve estimates of PVE were calculated as the \( R^2 \) from simple linear regression of the trait BLUP, including simulated additive effects, on SNV genotypes. This was only performed for SNVs that were significantly associated with the simulated trait BLUP at \( \alpha = 10^{-5} \).

Data and assigned constant additive effects \( -a, 0 \) and \( a \) (Falconer, 1989) to genotypes containing 0, 1 and 2 copies, respectively, of an arbitrarily chosen allele (all SNVs used in this study were bi-allelic). These effects were then added to phenotypic trait BLUPs, and the resulting data were analyzed using EMMAX, as described above. This approach preserves the overall structure of the data, whilst also allowing us to estimate statistical power across a range of allele effect sizes, as well as to compare estimated with expected PVE values. The expected value of PVE was calculated as:

---

© 2013 The Authors
New Phytologist © 2013 New Phytologist Trust

www.newphytologist.com
Results

RAD-Seq genotyping

Our genotyping approach resulted in the detection of over 100 000 putative SNVs (see the Materials and Methods section). A large proportion of these SNVs did not satisfy even our ‘liberal’ filtering criteria, but even after ‘stringent’ filtering, over 20 000 informative loci were available for downstream analyses (Table 2). As expected, the chromosome-wide distribution of SNVs appeared to be biased, with very few SNVs detected in putative centromeric regions (Fig. 3). At a finer scale, the vast majority of SNVs detected based on alignments to Sorghum (i.e. 98% and 96%, respectively, for ‘stringent’ and ‘liberal’ filtering) appeared to be located either inside or within 2 kb of genes.

As expected from the nature of RAD-Seq data, the majority of RAD-Seq SNVs were unlinked. For example, assuming microsynteny between Sorghum and Miscanthus, we used the sliding window LD-based pruning option of PLINK to calculate that 35 700 of the 53 174 SNVs detected based on alignments to the Miscanthus genome, analyses based on SNVs located on the same RAD tag suggested that fine-scale LD is substantial (average $r^2 = 0.37$ for SNVs with $MAF \geq 0.10$, Fig. S1). Furthermore, analyses assuming microsynteny between Sorghum and Miscanthus clearly demonstrated that significant LD (average $r^2 \geq 0.2$) extends to at least several hundred bp (Table 2) and up to at least 1 Mb for some pairs of SNVs (Fig. S1).

Population structure

Individual-based PCA based on SNVs filtered using ‘stringent’ criteria (Table 2) resulted in the identification of five significant eigenvectors ($P < 10^{-4}$ from tests based on the Tracy–Widom distribution), which explained between 2.9% and 1.2% of the total SNV variation. As expected, PCA (Fig. 1) and model-based clustering (Fig. S2) consistently delineated a ‘Continent’ vs ‘Japan’ genetic discontinuity, with both PC1 (Pearson’s $r = 0.83$, $P < 10^{-15}$) and PC2 (Pearson’s $r = 0.72$, $P < 10^{-11}$) being strongly correlated with the source longitudes of genotypes with known sampling locations. The differentiation between these two subpopulations was reflected in fine-scale patterns of LD, with $r^2$ decaying slightly more slowly in each subpopulation than in the overall population (Fig. S1). Furthermore, the clear pattern of clustering based on PC1 and PC2, which accounted for 2.9% and 2.0% of the SNV variation, enabled us to form strong hypotheses about the geographical origins of 68% (47 of 69) of the accessions with unknown sampling locations (Fig. 1).

GWASs

As expected from the relatively small size of our population, data perturbation simulations indicated that GWAS analyses had very limited power to detect associations with small or moderate effects ($PVE \leq 0.10$, Fig. S3). When such associations were detected, their estimated effects tended to be upwardly biased by as much as an order of magnitude (Fig. S3).

Naïve GWAS analyses that ignored the effects of population structure and relatedness (i.e. simple linear regression of trait BLUPs on individual SNVs) consistently resulted in severely inflated $P$ values (i.e. quantile–quantile (QQ) plots in Figs S4, S5). This effect was particularly strong for traits characterized by strong genetic differentiation (i.e. high $Q_{ST}$ between ‘Continent’ and ‘Japan’ subpopulations) and/or significant correlation with geographical variables or primary eigenvectors of population structure (Table 1, Figs S4, S5). The inclusion of the IBS kinship matrix substantially mitigated the confounding for all traits, whereas adding PC1 and PC2 further reduced the inflation of $P$ values for some traits (i.e. $BaseDiameter.9$, $LeafWidth.7$, $TransectCount.9$), without apparently compromising statistical
power relative to models that only included the IBS matrix (Figs S4, S5).

Using this conservative approach (i.e. including both the IBS kinship matrix and PC1 and PC2 as covariates in EMMAX analyses), we detected 35 putative associations ($P < 10^{-5}$) for SNVs resulting from alignments to the *S. bicolor* genome (Fig. S4). Four of these associations (two for *AvgeSen.9*, one for *LeafLength.7* and one for *Lignin.8*) reached genome-wide significance after Bonferroni correction for multiple testing ($P < 0.05/53174 \approx 9.4 \times 10^{-7}$), whereas another 13 had an estimated false discovery rate < 0.05 (Table 3). We detected similar patterns using SNVs resulting from alignments to an *M. sinensis* pseudo-reference (Fig. S5), with 53 SNVs reaching suggestive ($P < 10^{-5}$) and two SNVs (one for *DOYFS1.9* and one for *TallestStem.9*) reaching Bonferroni-corrected genome-wide significance ($P < 0.05/121771 \approx 4.1 \times 10^{-7}$).

Results from the MLMM-GWAS approach were generally consistent with those from the single-locus EMMAX analyses (Figs S4, S6). However, the MLMM approach allowed us to detect Bonferroni-corrected associations for two additional traits (*BaseDiameter.9* and *TallestStem.9*). Furthermore, when PC1 and PC2 were included as covariates in MLMMs, the selection procedure for *AvgeSen.9* favored a model including six significantly associated SNVs (Fig. S7).

**Genome-wide prediction**

As expected, predictive abilities were moderately correlated with broad-sense heritabilities (*Pearson’s r* > 0.57, *P* < 0.018, Table 4) and appeared to increase monotonically with the size of the training population (Fig. 4a). Although considerable variation was present among the 17 traits, the intercepts and slopes

---

**Table 4** Markers with significant phenotypic associations (false discovery rate < 0.05) in 138 Miscanthus sinensis genotypes

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>P^b</th>
<th>Q^d</th>
<th>MAF^b</th>
<th>PVE^f</th>
<th>Trait^g</th>
<th>Gene^h</th>
<th>Description/annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>377666</td>
<td>3.31E-06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.12</td>
<td>TallestStem.9</td>
<td>Sb01g004700</td>
<td>ATVPMP725</td>
</tr>
<tr>
<td>1</td>
<td>3789996</td>
<td>2.98E-06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.13</td>
<td>TallestStem.9</td>
<td>Sb01g004720</td>
<td>Aminoacyl-tRNA synthetase family</td>
</tr>
<tr>
<td>1</td>
<td>68013200</td>
<td>3.38E-06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.12</td>
<td>TallestStem.9</td>
<td>Sb01g044850</td>
<td>Unknown protein</td>
</tr>
<tr>
<td>2</td>
<td>67477249</td>
<td>1.14E-06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.16</td>
<td>StatusStrangle.7</td>
<td>Sb02g032850</td>
<td>Unknown protein</td>
</tr>
<tr>
<td>2</td>
<td>67477259</td>
<td>1.14E-06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.16</td>
<td>StatusStrangle.7</td>
<td>Sb02g032850</td>
<td>Unknown protein</td>
</tr>
<tr>
<td>3</td>
<td>8793225</td>
<td>6.38E-07</td>
<td>0.03</td>
<td>0.49</td>
<td>0.18</td>
<td>LeafLength.7</td>
<td>Sb03g008300</td>
<td>DNA binding/protein dimerization</td>
</tr>
<tr>
<td>3</td>
<td>65293238</td>
<td>2.79E-06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.02</td>
<td>AvgeSen.9</td>
<td>Sb03g037130</td>
<td>ATCPMEK</td>
</tr>
<tr>
<td>3</td>
<td>65293239</td>
<td>2.79E-06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.02</td>
<td>AvgeSen.9</td>
<td>Sb03g037130</td>
<td>ATCPMEK</td>
</tr>
<tr>
<td>4</td>
<td>15105860</td>
<td>1.57E-06</td>
<td>0.04</td>
<td>0.06</td>
<td>0.18</td>
<td>LeafLength.7</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>52494850</td>
<td>3.35E-06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.16</td>
<td>TallestStem.9</td>
<td>Sb06g020830</td>
<td>Protein kinase family protein</td>
</tr>
<tr>
<td>6</td>
<td>54211630</td>
<td>8.64E-07</td>
<td>0.05</td>
<td>0.17</td>
<td>0.15</td>
<td>Lignin.8</td>
<td>Sb06g025250</td>
<td>Serine-type endopeptidase/serine-type peptidase</td>
</tr>
<tr>
<td>6</td>
<td>58247799</td>
<td>5.66E-07</td>
<td>0.02</td>
<td>0.03</td>
<td>0.25</td>
<td>AvgeSen.9</td>
<td>Sb06g029670</td>
<td>ATP binding/protein kinase/protein serine/threonine kinase/protein tyrosine kinase/sugar binding</td>
</tr>
<tr>
<td>9</td>
<td>24591741</td>
<td>1.84E-06</td>
<td>0.03</td>
<td>0.03</td>
<td>0.26</td>
<td>AvgeSen.9</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>9</td>
<td>46480666</td>
<td>4.07E-06</td>
<td>0.04</td>
<td>0.15</td>
<td>0.02</td>
<td>AvgeSen.9</td>
<td>Sb09g018620</td>
<td>Hydroxyproline-rich glycoprotein family protein</td>
</tr>
<tr>
<td>10</td>
<td>49562793</td>
<td>4.90E-07</td>
<td>0.02</td>
<td>0.04</td>
<td>0.21</td>
<td>AvgeSen.9</td>
<td>Sb10g022360</td>
<td>Unknown protein</td>
</tr>
<tr>
<td>10</td>
<td>55335850</td>
<td>3.14E-06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.12</td>
<td>TallestStem.9</td>
<td>Sb10g026010</td>
<td>UBP19; cysteine-type endopeptidase/ubiquitin thiolesterase</td>
</tr>
<tr>
<td>10</td>
<td>59574096</td>
<td>1.50E-06</td>
<td>0.03</td>
<td>0.05</td>
<td>0.19</td>
<td>TallestStem.9</td>
<td>Sb10g029835</td>
<td>Unknown protein</td>
</tr>
</tbody>
</table>

Associations with Bonferroni-corrected genome-wide significance ($P < 0.05$) are shown in bold. Only results for markers detected from alignments to the *Sorghum bicolor* genome are shown.

^aChromosome, *Sorghum bicolor* chromosome to which the marker was aligned.

^bPosition, *Sorghum bicolor* chromosome position to which the marker was aligned.

^cP, *P* value from Genome-wide association studies (GWAS) analysis using the efficient mixed-model association expedited approach (EMMAX), including the kinship matrix and the first two eigenvectors of population structure (see the Materials and Methods section).

^dQ, false discovery rate calculated using the *q*-value R package (Dabney & Storey, 2013).

^eMAF, minor allele frequency.

^fPVE, naive estimate of the proportion of variance explained based on simple linear regression (see the Materials and Methods section).

^gTrait, phenotypic trait as defined in Table 1.

^hGene, *Sorghum bicolor* gene to which the marker was aligned.

^iSignificant at genome-wide level after Bonferroni correction based on EMMAX analyses (see the Materials and Methods section).

^jIncluded in the optimal model according to the multiple Bonferroni criterion in multi-locus mixed-model (MLMM) analyses (see the Materials and Methods section).

^kIncluded in the optimal model according to the multiple Bonferroni criterion in MLMM analyses including the first two eigenvectors of population structure (see the Materials and Methods section).

NA, not applicable (markers aligning to putatively intergenic positions).
of simple linear regressions of BLUPs calculated from field data on those estimated using ridge regression tended to be close to their expected values of 0.00 and 1.00, respectively (Table S1). Interestingly, both predictive abilities and accuracies of genome-wide prediction were statistically indistinguishable (P > 0.64 from paired t-tests across traits) between the sets of SNVs obtained based on different alignments (Table 4). Furthermore, predictive abilities tended to reach a plateau when the performance of genome-wide prediction, with predictive abilities from cross-validations across subpopulations being significantly lower than those from corresponding random cross-validations (one-sided P = 0.0004 from a paired t-test across traits). However, the extent of this difference varied considerably among traits (Fig. 5). Finally, the selection of markers based on their GWAS significance or rrBLUP-estimated effects appears to have potential for improving predictive abilities (Fig. 6), particularly for traits with lower heritabilities (Fig. 6b).

Discussion

As in a number of previous studies across a wide phylogenetic range of plant species (Barchi et al., 2011; Chutimanitsakun et al., 2011; Stötting et al., 2013), the application of RAD-Seq genotyping resulted in the generation of large numbers of informative SNV markers (Table 2). Alignments of RAD tags to an M. sinensis pseudo-reference resulted in the identification of substantially greater numbers of markers (Table 2) and indistinguishable performance of genome-wide prediction (Table 4) compared with alignments to the S. bicolor genome. However, it is possible that the availability of a reference genome sequence from a closer relative would have resulted in more informative RAD-Seq SNV data. In any case, our results indicate that the a priori availability of a high-quality reference genome sequence does not appear to be a requirement for the success of this genotyping procedure.

Although we were able to generate data for a large number of RAD-Seq markers, the use of a methylation-sensitive enzyme (PstI) and relatively stringent alignment criteria resulted in a greatly unbalanced genome coverage (Fig. 3) and strong bias against intergenic SNVs. Although probably advantageous in
terms of data quality, this effect may limit the applicability of RAD-Seq genotyping for the complete dissection of complex trait architecture because over half of trait-associated polymorphisms may be located outside of genes (Li et al., 2012). However, this shortcoming could possibly be mitigated by using multiple restriction enzymes with varying sensitivities to methylation.

The patterns of putatively neutral population structure detected based on RAD-Seq SNVs (Fig. 1) were consistent with those described previously based on much smaller numbers of markers (Slavov et al., 2013b). However, the greater power of the RAD-Seq markers allowed us to detect additional nuances, as well as to form hypotheses about the geographical origins of over two-thirds of the genotypes with unknown sampling locations (Fig. 1). Extensions of this approach to wider germplasm collections (including other Miscanthus species) and more sophisticated models (Baran et al., 2013) will significantly improve our knowledge about the evolutionary history of the genus and will provide important practical information to breeders.

Despite the small size of our association mapping population (N = 138), seven associations consistently reached genome-wide significance across a range of GWAS analyses and adjustments for multiple testing (Table 3, Figs S4–S7). Interestingly, two of these associations (i.e. one for AvgeSen.9 and one for Lignin.8; Table 3), as well as another two putative associations for StemDiameter.9 and TallestStem.9 (P < 10⁻⁵, Fig. S4), all appeared to align within 8–16 Mb of the putative dwarfing locus dw2, whose location was recently supported by a GWAS for plant height in S. bicolor (Morris et al., 2013). This region may therefore require particular attention in future association and linkage mapping studies.

What is the explanation for the detection of multiple significant associations, given the relatively limited statistical power of our GWASs (Fig. S3)? Because no single unambiguous explanation can be provided, we hypothesize that several factors may underlie this observation. First, LD in M. sinensis (Fig. S1) does not appear to be as extensive as in its primarily self-pollinating
relative *S. bicolor* (Morris et al., 2013). However, there was clear evidence of long-range LD in at least some regions of the *Miscanthus* genome (Fig. S1). Thus, although the initial stages of follow-up studies will focus on the regions in the immediate vicinity of trait-associated SNVs, it is conceivable that at least some of these are tagging causative polymorphisms located many kilobases away. Second, the statistical power to detect associations of small to moderate effect (PVE $\leq 0.10$) was only between 0.001 and 0.213 (Fig. S3). However, if the genomic architectures of the traits that we studied are highly complex, dozens or even hundreds of causative polymorphisms with minor effects may exist across the genome. Under this highly polygenic scenario, even an underpowered study that uses a large number of markers is likely to detect a subset of minor effect associations and dramatically overestimate their PVE (Fig. S3). Alternatively, some of the SNVs detected by our GWASs may be linked to causative polymorphisms of larger effects. However, the former scenario seems to be more plausible based on larger scale GWAS results in other related crops (Buckler et al., 2009; Morris et al., 2013). Furthermore, the highly polygenic scenario is consistent with the results from our assessment of marker selection strategies for genome-wide prediction (Fig. 6), which was performed through cross-validation and should therefore be immune to the inflation of effect sizes. Finally, it is possible that our analytical procedures failed to control the rate of false positives. However, our results (i.e. QQ plots in Figs S4, S5) are generally inconsistent with this explanation.

Although GWAS results are expected to improve in the near future (i.e. with the use of larger populations and denser genome coverage), the ability to predict phenotypes from a genome-wide set of markers is likely to have immediate impact on *Miscanthus* breeding programs. Our genome-wide prediction results illustrate several important points. First, as expected from both theoretical and empirical studies (Daetwyler et al., 2008, 2010, 2013; Resende et al., 2012), trait heritability was correlated with predictive ability. However, this correlation was only moderate, confirming that other factors (e.g. genomic architecture of the trait, LD, effective population size) may be equally important (Daetwyler et al., 2010; de Los Campos et al., 2013). For example, *DryMatter.9* was moderately heritable ($H^2 = 0.54$), yet genome-wide prediction of this trait consistently failed (Table 4, Fig. 4). This is not surprising, given the highly composite nature of biomass yield, and targeting individual yield components or correlates (e.g. *MaxCanopyHeight.9* and *StemDiameter.9*, Fig. 2) appears to be a more promising approach (Table 4, Fig. 4). In contrast, the accuracy of genome-wide prediction for *Moisture.9* ($H^2 = 0.59$) nearly reached its theoretical maximum (Table 4), suggesting that genomic selection for this trait may be feasible even with very small training populations (i.e. $N < 100$, Fig. 4a), with as few as 100–1000 markers (Fig. 4b) and across subpopulations (Fig. 5). Predictive ability was similarly high for phenological traits (i.e. *DOYFS1.9* and *AgeSen.9*), as well as *Cellulose.8*, and promising for most other traits (i.e. except *DryMatter.9* and *TransectCount.9*, Table 4, Fig. 4). We are therefore exploring the practical application and validation of genomic selection in the *Miscanthus* breeding program at the Institute of Biological, Environmental and Rural Sciences (IBERS). Second, our results clearly suggest that substantial further improvements in predictive ability are more likely to come from using larger training populations (Fig. 4a) than from using denser genome coverage (Fig. 4b). However, an important caveat is that the SNVs that we used probably did not cover vast regions of the *Miscanthus* genome (Fig. 3). The observed plateau in predictive ability when 10 000–20 000 SNVs were used clearly needs to be validated using more representative samples of markers. Based on our results and currently available genotyping platforms, our

---

**Fig. 6** Effect of marker selection on the performance of genome-wide prediction of average senescence score (a) and total dry weight (b) in a population of 138 *Miscanthus sinensis* genotypes based on single-nucleotide variants (SNVs) detected from alignments to the *Sorghum bicolor* genome. ‘Random’ (circles, black line), randomly selected markers; ‘GWAS’ (triangles, red line), markers with the lowest genome-wide association study $P$ values within the training population; ‘rrBLUP’ (squares, green line), markers with the highest estimated effects within the training population based on ridge regression. All data points are averages ± SD across 100 random cross-validations. Dashed lines correspond to predictive abilities based on all 53 174 SNVs.
recommended approach to genome-wide prediction in non-
model plants would be to maximize the number of individuals in
the training population and to use a low-cost genotyping strategy
(Davey et al., 2011; Elshire et al., 2011; Poland et al., 2012).
Third, the robustness of genome-wide prediction across subpop-
ulations varied dramatically among traits (Fig. 5). Thus, the rela-
tionship between training and test populations needs to be
characterized in detail, and individuals from the populations tar-
geted by genomic selection (or their close relatives) should ideally
be included in training populations. Finally, our preliminary
assessment of locus selection strategies (Fig. 6) clearly indicated
that there is great potential to increase predictive abilities through
the application of sophisticated analytical approaches (de Los
Campos et al., 2013), which do not attempt to estimate pheno-
typic effects for all loci (i.e. in contrast to the method we used).
The refinement of these approaches and their combination with
multi-locus and multi-trait procedures (Korte et al., 2012; Segura
et al., 2012) offer exciting prospects for the characterization of
pleiotropy and the dissection of highly complex phenotypic traits.
However, studies aimed at pushing the limits of genome-wide
prediction cannot overcome the inherent limitations of this approach (e.g. trait heritability sets an upper limit on accuracy)
and need to be designed and interpreted with awareness of its
numerous pitfalls (Wray et al., 2013).

In summary, we related high-quality phenotypic data for 17
traits in a population of *M. sinensis* to SNV markers obtained
through RAD-Seq genotyping. Despite the relatively small size of
our experimental population (N=138), results from GWASs were
promising and suggest that this approach will be instrumenta-
lar for the dissection of complex phenotypic traits. On a more
immediate time-scale, results from our genome-wide prediction
analyses suggest that the application of genomic selection in
*Miscanthus* may be feasible, and we are therefore validating this
finding in our accelerated breeding program.

Acknowledgements

G.S. was funded by the Biosciences, Environment and Agricultu-
re Alliance whilst working on the manuscript. Funding was also
provided by the UK Biotechnology and Biological Sciences
Research Council (BBSRC) Institute Strategic Programme grant
BB/J0042/1 and by BBSRC responsive mode grant BB/
K01711X/1. We thank the Editor and three anonymous review-
ers for their comments on an earlier version of the manuscript, as
well as John Norris for providing Linux computing support.

References

Allison DB, Fernandez JR, Heo M, Zhu SK, Etzel C, Beasley TM, Amos CI.
2002. Bias in estimates of quantitative-trait-locus effect in genome
scans: demonstration of the phenomenon and a method-of-moments
Genotypic variation in cell wall composition in a diverse set of 244 accessions

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU,
Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping
localization of genetic samples through linkage-disequilibrium correction.
Barchi L, Lanteri S, Portis E, Aquadro A, Vale G, Toppino L, Rotino GL.
2011. Identification of SNP and SSR markers in eggplant using RAD tag
AH, ed. *Molecular dissection of complex traits*. Boca Raton, FL, USA: CRC
Press, 145–162.
Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistue L, Corey A,
Filichkina T, Johnson EA, Hayes PM. 2011. Construction and application for
QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in
Clifton-Brown JC, Breuer J, Jones MB. 2007. Carbon mitigation by the energy
for energy in Europe and its potential contribution to decreasing fossil fuel
Dahney A, Storey JD. 2013. qvalue: Q-value estimation for false discovery rate
control. R package version 1.34.0. URL: http://www.bioconductor.org/packages/
Daetwyler HD, Calus MPL, Wong-Wong R, de Los Campos G, Hickey JY.
2013. Genomic prediction in animals and plants: simulation of data, validation,
genetic architecture on genome-wide evaluation methods. *Genetics* 185:
1021–1031.
genetic risk of disease using a genome-wide approach. *PLoS One* 3:
e3395.
Danecik P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA,
Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call
Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML.
2011. Genome-wide genetic marker discovery and genotyping using
Earl D, vonHoldt B. 2012. STRUCTURE HARVESTER: a website and
program for visualizing STRUCTURE output and implementing the Evanno
Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell
SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for
Endelman JB. 2011. Ridge regression and other kernels for genomic selection
with R package rrBLUP. *Plant Genome* 4: 250–255.
Evanno G, Regnaut S, Goudet J. 2005. Inference of population structure
using multilocus genotype data: linked loci and correlated allele frequencies.
Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure
using multilocus genotype data: linked loci and correlated allele frequencies.
*Genetics* 164: 1567–1587.
Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure
using multilocus genotype data: dominant markers and null alleles. *Molecular
Ecology Notes* 7: 574–578.
Feltus FA, Vandenbergk JP. 2012. Bioenergy grass feedstock: current options and
prospects for trait improvement using emerging genetic, genomic, and systems


Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81 : 559–575.


Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Linkage disequilibrium based on single-nucleotide variants detected using restriction site-associated DNA sequencing in Miscanthus sinensis.

Fig. S2 Model-based clustering of single-nucleotide variant data for 138 Miscanthus sinensis genotypes.
**Fig. S3** Statistical power and effect size inflation based on data perturbation simulations for genome-wide association study analyses using the EMMAX program.

**Fig. S4** Genome-wide association study (GWAS) results for 17 phenotypic traits in a population of 138 Miscanthus sinensis genotypes based on 53 174 markers detected using alignments to the Sorghum bicolor genome.

**Fig. S5** Genome-wide association study (GWAS) results for 17 phenotypic traits in a population of 138 Miscanthus sinensis genotypes based on 121 771 markers detected using alignments to a M. sinensis pseudo-reference.

**Fig. S6** Multi-locus mixed-model (MLMM) genome-wide association study results for 17 phenotypic traits in a population of 138 Miscanthus sinensis genotypes based on 53 174 markers detected using alignments to the Sorghum bicolor genome.

**Fig. S7** Covariate multi-locus mixed-model (MLMM) genome-wide association study (GWAS) results for average senescence score in a population of 138 Miscanthus sinensis genotypes based on 53 174 markers detected using alignments to the Sorghum bicolor genome.

**Fig. S8** Genetic relationship matrix among 138 Miscanthus sinensis genotypes.

**Table S1** Simple linear regression measures of performance of genome-wide prediction in a population of 138 Miscanthus sinensis genotypes.

**Methods S1** Miscanthus pseudo-reference.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the New Phytologist Central Office.