2017

Breeding for Biomass Yield in Switchgrass Using Surrogate Measures of Yield

Michael D. Casler
USDA-ARS, michael.casler@ars.usda.gov

Guillaume P. Ramstein
University of Wisconsin- Madison, ramstein@wisc.edu

Bioenerg. Res.
DOI 10.1007/s12155-017-9867-y

Breeding for Biomass Yield in Switchgrass Using Surrogate Measures of Yield

Michael D. Casler[1] · **Guillaume P. Ramstein**[2]

Abstract Development of switchgrass (*Panicum virgatum* L.) as a dedicated biomass crop for conversion to energy requires substantial increases in biomass yield. Most efforts to breed for increased biomass yield are based on some form of indirect selection. The objective of this paper is to evaluate and compare the expected efficiency of several indirect measures of breeding value for improving sward-plot biomass yield of switchgrass. Sward-plot biomass yield, row-plot biomass, and spaced-plant biomass were measured on 144 half-sib families or their maternal parents from the WS4U-C2 breeding population of upland switchgrass. Heading date was also scored on row plots and anthesis date was scored on spaced plants. Use of any of these indirect selection criteria was expected to be less efficient than direct selection for biomass yield measured on sward plots, when expressed as genetic gain per year. Combining any of these indirect selection criteria with half-sib family selection for biomass yield resulted in increases in efficiency of 14 to 36%, but this could only be achieved at a very large cost of measuring phenotype on literally thousands of plants that would eventually have no chance of being selected because they were derived from inferior families. Genomic prediction methods offered the best solution to increase breeding efficiency by reducing average cycle time, increasing selection intensity, and placing selection pressure on all additive genetic variance within the population. Use of genomic selection methods is expected to double or triple genetic gains over field-based half-sib family selection.

Keywords *Panicum virgatum* · Genomic selection · Genomic prediction · Biomass yield

Introduction

Development of switchgrass (*Panicum virgatum* L.) as a dedicated biomass crop for conversion to energy requires substantial increases in biomass yield. Using the best available forage-type cultivars, biomass production can be economic and sustainable only under the best management conditions, which are generally more likely to be achieved by the most experienced growers [1]. Biomass yield is a heritable trait in switchgrass and yield gains have been achieved in several breeding programs during the past 20 years [2–4]. Gains in biomass yield as high as 27% have been reported from a single generation of selection within wild or relatively unimproved populations, e.g., [3]. However, long-term gains from sustained selection and breeding are in the range of 1 to 4% year^{-1} for both upland and lowland ecotypes of switchgrass [2].

Breeding for increased biomass yield in switchgrass presents the breeder with the fundamental challenge of measuring biomass yield accurately and precisely. Selection of individual widely spaced plants using classical phenotypic selection methods can only go so far. Spaced-plant biomass is not equivalent to sward-plot biomass yield—these are different traits with a highly variable and inconsistent genetic correlation [5]. At the most fundamental level, spaced plantings or row plantings do not allow for any interplant competition,

Electronic supplementary material The online version of this article (doi:10.1007/s12155-017-9867-y) contains supplementary material, which is available to authorized users.

✉ Michael D. Casler
michael.casler@ars.usda.gov; mdcasler@wisc.edu

[1] USDA-ARS, US Dairy Forage Res. Center, 1925 Linden Dr. West, Madison, WI 53706-1108, USA

[2] Department of Agronomy, University of Wisconsin, Madison, WI 53706, USA

Published online: 10 September 2017

whereas sward-plot likely include a significant mortality factor due to intensive interplant competition. Phenotypic selection of individual plants with the greatest amount of biomass is very effective on relatively raw and unimproved germplasm [2–4], easily allowing breeders to discard plants with low vigor, tillering, or whole-plant biomass. However, the genetic correlation between spaced-plant biomass and sward-plot biomass yield diminishes to near zero for elite and highly improved germplasm [6]. Any use of sward-plot evaluation increases both time and funding costs. Clonal propagation of individual genotypes is prohibitive in switchgrass, so sward plot evaluations require seed produced from either full-sib or half-sib matings. Usually, these matings must occur in the field in order to generate sufficient seed for field trials at a minimum of two locations [7]. All of these are reasons that sward-plot field trials are seldom used in switchgrass breeding.

Indirect selection is a potential solution to these challenges. Strong genetic correlations have been reported for spaced-plant biomass with morphological or physiological traits such as tiller number, tiller mass, and flowering time [8, 9]. These correlations imply that these traits could be used as indirect selection criteria to either speed up the breeding process or to increase selection intensity per generation, potentially increasing the rate of gain for biomass yield. While there is some evidence that these potential indirect selection criteria may have positive genetic correlations with sward-plot biomass yield [10], the only results to date have demonstrated a near-zero genetic correlation for all vigor-related or tillering traits [6]. Indeed, divergent selection for several of these traits failed to generate any correlated responses in sward-plot biomass yield [11]. Conversely, selection of late flowering genotypes in spaced plantings have resulted in strong correlated responses for increased biomass yield of sward plots [2, 11].

Genomic prediction is an alternative approach to accomplish indirect selection of biomass yield, involving the use of genome-wide DNA markers to predict breeding value of elite genotypes [12]. Selection based on genomic estimated breeding values (GEBV) allows selection pressure to be placed on a large number of loci impacting the trait of interest, limited only by genomic marker coverage. Selection based on GEBV requires a training period, in which prediction equations are developed to predict GEBV of unknown genotypes using observed relationships between markers and biomass yield [13]. The advantage of selection on GEBV derives from the next step, which involves two or three generations of rapid selection on GEBV, which can theoretically be accomplished in 1 year per generation.

The objective of this paper is to evaluate and compare the expected efficiency of several indirect measures of breeding value for improving sward-plot biomass yield of switchgrass. Both phenotypic (field-based) predictors and GEBV predictors were investigated in this study.

Materials and Methods

Plant Materials, Field Designs, and Field Data

A total of 144 half-sib families were developed from the WS4U-C2 switchgrass population, as previously described [5, 11]. Parental genotypes were vegetatively propagated and established in a randomized complete block design with three replicates at Arlington, WI (Plano silt loam; fine-silty, mixed, mesic Typic Argiudoll). The parental experiment was established in 2009 using vegetative propagules with approximately 8–12 tillers each.

Half-sib families were established in two types of plots: row plots and sward plots. Row plots consisted of five 12-week old seedlings transplanted to the field in May 2011, with spacing of 0.3 m within rows and 0.9 m between rows. The design was a randomized complete block with four replicates at each of two locations: Arlington, WI and Mead, NE (Haynie very fine sandy loam; coarse-silty, mixed, superactive, calcareous, mesic Mollic Udifluvent). Sward plots consisted of five drilled rows, 15 cm apart, with a seeding rate of 600 PLS m^{-2}. Sward plots were 0.9 m wide and 1.8 m long, and established at Arlington, WI and Marshfield, WI (Withee silt loam; fine-loamy, mixed, superactive frigid Aquic Glossudalf) in May 2008. The design was a randomized complete block with three replicates at each location. All plants and plots were allowed to grow during the establishment year and biomass was removed after killing frost. No fertilizer was applied during the establishment year. Pre-emergence herbicide was applied to the parental experiment and the progeny row plots before establishment as follows: application of 1.12 kg ha^{-1} alachlor [2-chloro-*N*-2,6-diethylphenyl)-*N*-(methoxymethyl)-acetamide] with 0.07 kg ha^{-1} imazethapyr {(±)-2-[4,5-dihydro-4-methyl-4-(1-methylethyl)-5-oxo-1*H*-imidazol-2-yl]-5-ethyl-3-pyridine- carboxylic acid}.

All plants and plots were fertilized with 112 kg N ha^{-1} in early spring following the establishment year. Pre-emergence herbicide was applied to the parental experiment and the progeny row plots as described above. Biomass from all plants and plots was harvested between 2 and 4 weeks post-anthesis at a cutting height of 9 cm. Parental genotypes were cut with a sickle-bar mower and weighed by hand. Progeny row plots and sward plots were cut with a flail chopper and plot weights determined by a load cell. Biomass samples of approximately 200 to 400 g were sampled from each plant or plot, dried at 60 °C for 5–7 days, and reweighed to determine dry matter concentration. All biomass yields were adjusted to a dry matter basis. Biomass yields were determined for 4 years on sward plots (2008–2011) and 2 years each for row plots (2012–2013) and parental spaced plants (2008–2009).

Heading date of progeny row plots was determined at both locations for 2 years. Heading date was scored on each

individual plant in every row and defined as the calendar date on which approximately 50% of the panicles were fully emerged from the boot. Anthesis date was determined on parental spaced plants for 2 years, defined as the calendar date on which approximately 50% of the panicles reached anthesis [6].

Biomass Yield and Field Traits

Sward-plot biomass yield, measured on half-sib family plots for 4 years at Arlington and Marshfield, was used as the best estimate of the true breeding value of these 144 parental genotypes. Best linear unbiased predictors (BLUPs) were used as estimates of the breeding values for Arlington, Marshfield, and means over two locations, depending on the specific analysis described below.

Spaced-plant biomass of the maternal parents and row-plot biomass of the half-sib families were used as two different indirect measures of breeding value. Spaced plantings are commonly used in breeding perennial grasses, but are characterized by a complete lack of competition between neighbors. Row plots have been proposed as a compromise, creating an opportunity to evaluate breeding value of genotypes by testing their half-sib progeny in a plot type that contains a modest level of competition between neighbors. Heading date of parental spaced plants and anthesis date of progeny row plots were also used as indirect predictors of sward-plot biomass yield on a per-area basis.

General linear mixed models were used to estimate variance components for all effects. There were no fixed effects in any of these models. Heritability was computed as the ratio of estimated additive genetic variance to phenotypic variance, except for parental spaced plants, where the heritability could be computed only in the broad sense as the ratio of genotypic to phenotypic variance. Genetic correlations were estimated as the correlation between BLUPs for each trait; this was the only mechanism of estimating the genetic correlation between observed biomass yield and GEBV, because GEBVs consisted of a single vector of values, one per family [14].

Expected gains from selection were computed as the ratio of correlated response ($CR_{Y|X}$) trait Y (sward-plot biomass yield) from indirect selection for trait X. All values of $CR_{Y|X}$ were expressed as a percentage of the expected direct response (R_Y) for selection based on sward-plot biomass yield. Three selection schemes were used to generate expectations: (1) half-sib family selection with intercrossing of parental genotypes (HSF) with 5 years per cycle, (2) individual phenotypic selection (PS) with 2 years per cycle, and (3) combined half-sib family selection for biomass yield and within-family indirect selection for trait X (AWFX-HS) with 4 years per cycle [15]. For HSF and PS, $CR_{Y|X}/R_Y$ computations were made according to equation 19.9 of [16]. For AWFX-HS, $CR_{Y|X}/R_Y$ computations were made as the ratio of equation 4 to equation 1 of [15].

Genomic DNA Data and Prediction Methods

Marker Sequencing, Genotyping, and Imputation

DNA marker data were used to develop genomic prediction equations for sward-plot biomass yield. DNA markers were single nucleotide polymorphisms (SNPs) scored on maternal parents. Biomass yield was measured on half-sib families. Two methods for generating sequence data were assessed: exome-capture sequencing (ECS), with targeted coverage and high sequencing depth, and genotyping-by-sequencing (GBS), with (generally) broader genome coverage but lower sequencing depth. When calling SNPs, the sequences generated by either method were aligned to the hardmasked *P. virgatum* v1.1 reference genome (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum).

In ECS, specific exonic sequences were captured using the Roche-Nimblegen protocol for preparation of SeqCap EZ Developer libraries using the Roche-Nimblegen probeset '120911_Switchrass_GLBRC_R_EZ_HX1' [17, 18]. Sequencing was performed on the Illumina HiSeq 2000 platform, generating 150-nucleotide paired-end reads. In GBS, the genome space was reduced using restriction enzymes PstI and MspI, following [19]. Sequencing was barcoded and multiplexed 96 times. The Illumina HiSeq 2500 platform was used to generate 100-nucleotide single-end reads.

In ECS, initial quality control (using FastQC v0.10.0; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), sequence trimming (using Cutadapt v1.1; https://code.google.com/p/cutadapt/), alignment using BowTie v0.12.7 [20], SNP calling using SAMTools package v0.1.18 [21], and genotype calling using the algorithm of [22], as shown in the R script in File S1, were performed as described in [13].

In GBS, initial quality control was performed using FastQC v0.11.2. Adapters were trimmed using Skewer [23] and the resulting sequences reads were merged into tags, with counts, using TASSEL v4.3.6 [24, 25]. Tags were filtered for counts greater than 10 and merged into a master tag count file. The master tags were then aligned to the reference genome using Bowtie v2.2.1 [26], to generate a "Tags On Physical Map" (TOPM) file. Tag counts by individual (maternal parent) were determined from barcodes and stored in a "Tags by Taxa" (TBT) file. The TOPM and TBT files were subsequently used for calling SNPs. Genotype were then called directly from allele counts, with no account of possible false homozygote calls.

In ECS and GBS, marker variables from the matrices of genotype calls were filtered for (i) proportion of missing values (strictly lower than 5% in ECS, strictly lower than 80% in GBS), (ii) polymorphism (variance strictly greater than 0 and minor allele frequency strictly greater than $1/2N$,

with N the number of genotypes assayed), and (iii) availability of genomic-location information (available information on chromosome and position from the reference genome sequence and annotation of *P. virgatum* v1.1; DOE-JGI, http://phytozome.jgi.doe.gov/). The resulting marker genotype matrices $\mathbf{M}_{ECS}$ and $\mathbf{M}_{GBS}$, obtained from ECS and GBS, contained expected allelic dosages at $q^* = 120{,}203$ markers and hard genotype calls at $q^* = 10{,}856$ markers, respectively.

In ECS, missing values were imputed using the multivariate normal expectation-maximization algorithm of [27], implemented in the R package rrBLUP [28]. In GBS, missing values were imputed using three different approaches: (i) mean imputation (MI); (ii) imputation by hidden Markov model (HMM), as implemented in Beagle v4.1 [29]; and (iii) imputation by iterated Random Forest (RFI), as implemented in the R package missForest [30]. When imputing by MI, missing values were replaced with the average allelic dosage at each marker. When imputing by HMM, Beagle v4.1 was run over 20 iterations, preceded by 10 burn-in iterations, with an assumed recombination rate of 1 cM/Mb, as was suggested by a previous mapping study on switchgrass [31]; window size was set to 48 markers, with an overlap of 7 markers between windows, based on the developers' recommendation of about 5 cM per window with a 1-cM overlap and the pattern of linkage-disequilibrium decay in this population [13]; effective population size was set to 75, based on the observed heterozygote excess in the ECS data [32], which was deemed plausible given that WS4U-C2 was produced by two cycles of selection from a collection of 162 plants; error rate was set to 0.001. When imputing by RFI, Random Forest was run over 10 iterations, with 100 trees per iteration, bootstrap samples of size N and random subsets of 3618 markers per tree.

GEBV Predictions

Following the methodology described in [13], prediction procedures were evaluated with respect to three components: (1) marker-data transformation—potentially accounting for correlation among markers; (2) prediction model—potentially accounting for differential amplitudes (heteroscedasticity) and/or non-linearity of marker effects; and (3) environment learning scheme—set of locations to include for training and testing.

Four different prediction models were assessed: Genomic BLUP (GBLUP), a linear and homoscedastic model, i.e., with linear marker effects of equal variance [33, 34]; BayesB, a linear heteroscedastic model [35]; Reproducing kernel Hilbert space (RKHS), a non-linear homoscedastic model [36]; and Random Forests (RF), a non-linear heteroscedastic model [37].

The GBLUP model was considered the standard in model comparisons. For a sample of n instances and q marker features, we define GBLUP as follows:

$$\mathbf{g} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{g} = \{g_i\}$ is the n vector of HS-family BLUPs; $\boldsymbol{\mu}$ is the n vector of grand mean; $\mathbf{Z}$ is the $n \times m$ design matrix attributing the n observations to m parent genotypes; $\mathbf{u} \sim Normal(0, \mathbf{K}\sigma_u^2)$, $\mathbf{K}$ being the $m \times m$ genomic relationship matrix derived from marker features as $\mathbf{K} \propto \mathbf{XX}^{\mathbf{T}}$, with $\mathbf{X}$ the $m \times q$ matrix of marker features; $\mathbf{e} \sim Normal(0, \mathbf{I}\sigma_e^2)$, with $\mathbf{I}$ the identity matrix. As explained in the next subsection, the marker features in $\mathbf{X}$ were not necessarily the original marker variables, i.e., $\mathbf{X} \neq \mathbf{M}$. The normalizing factor in $\mathbf{K}$ was the sum of sample variances over marker features.

BayesB is a Bayesian linear regression model, which has the following specification:

$$\mathbf{g} = \boldsymbol{\mu} + \mathbf{Xb} + \mathbf{e}$$

where $\mathbf{g}$, $\boldsymbol{\mu}$, $\mathbf{X}$ are as described above; $\mathbf{b} \sim Normal(0, \mathbf{I}\sigma_b^2)$; $\sigma_b^2 = 0$ with probability π and $\sigma_b^2 \sim \chi^{-2}(df_b, S_b^2)$ with probability $1 - \pi$; π was chosen to follow a Beta(0.2, 1.8) in order to reflect possibly sparse distributions of causal variants across the genome while allowing uncertainty about π; $S_b^2 \sim Gamma(r_b, s_b)$ and $\mathbf{e} \sim Normal(0, \mathbf{I}\sigma_e^2)$, with $\sigma_e^2 \sim \chi^{-2}(df_e, S_e^2)$. The hyperparameters df_b, r_b, s_b, df_e and S_e^2 were set through the heuristics described in [38], based on a prior estimation of the proportion of variance explained by the model, which was here chosen to be $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$ from a GBLUP model with an update on marker effects from the heteroscedastic effects model (HEM) of [39]. BayesB was fitted by a Gibbs sampling algorithm with 5000 burn-in iterations, followed by 15,000 iterations for actual sampling of parameter values.

The RKHS is a semi-parametric model modeling relationships among individuals through a non-linear function of marker-based distances. The pairwise distances were Euclidean distances based on marker features, scaled by the maximum distance over pairs of individuals; the non-linear function was the Gaussian kernel, with its scale parameter determined by tuning, through minimization of the generalized-cross-validation criterion (GCV) [40, 41] over a grid of values (strictly greater than 0 and lower than 1, with steps of 0.025).

The RF model is a machine-learning method that combines results from several regression trees fitted to different variations of the data—bootstrap samples of instances and random subsets of features [37]. The RF model was fitted with 200 trees, bootstrap samples of size n, and subsets of $q/3$ features.

The GBLUP and RKHS models were fitted using the R package rrBLUP [28]; the BayesB models were fitted using the R package BGLR [36]; the HEM of was fitted using the R

package bigRR [39], and the RF model was fitted using the R package randomForest [42].

Marker-Data Transformations

The input **X** to prediction models were transformations of the marker-data matrix **M**. The following transformations of **M** were made: (1) *Base*, where features are centered allelic dosages: $\mathbf{X}_{Base} = \mathbf{M} - \mathbf{P}$, with **P** the $m \times q$ matrix with uniform columns containing the mean allelic dosages within the population, $q = q^*$; (2) *PCA*, where features are uncorrelated synthetic variables each contributed differentially by marker variables: $\mathbf{X}_{PCA} = \mathbf{X}_{Base}\mathbf{V}$, with **V** the $q^* \times d$ rotation matrix in the singular value decomposition of $\mathbf{X}_{Base}$ ($\mathbf{X}_{Base} = \mathbf{UDV}^{\mathrm{T}}$; d is the number of principal components, here equal to m), $q = d$; (3) *Cor*, where features are marker variables scaled through a correlation matrix: $\mathbf{X}_{Cor} = \mathbf{X}_{Base}\mathbf{R}^{-1/2}$, **R** being the block-diagonal matrix of Pearson correlation between marker variables (with blocks corresponding to chromosomes), and $\mathbf{R}^{-1/2}$ being the square-root, from eigendecomposition, of its inverse, $q = q^*$; (iv) *LD*, where features are marker variables weighted based on their relative degree of tagging: $\mathbf{X}_{LD} = \mathbf{X}_{Base}\mathbf{W}^{1/2}$, with **W** the diagonal matrix of weights supposed to adjust for redundancy in marker information due to linkage disequilibrium; $\mathbf{W} = \mathrm{diag}(\mathbf{w})$ and **w** was the least-absolute-error solution to $(\mathbf{R}\circ\mathbf{R})\,\mathbf{w} = 1_{q^*}$ subject to $w_j \geq 0, j = 1, \ldots, q^*$, with $\mathbf{R}^{\circ}\mathbf{R}$ the matrix of squared correlation between marker variables and 1_{q^*} the q^* vector of one values, $q \leq q^*$ (as a result of some weights being exactly zero). The linear programming solver CLP (https://projects.coin-or.org/Clp; R script in File S1) was used to calculate **w**. For computational tractability, when solving $(\mathbf{R}\circ\mathbf{R})\,\mathbf{w} = 1_{q^*}$, we applied two heuristics, following [43]: (i) values in $\mathbf{R}^{\circ}\mathbf{R}$ less than 0.001 were set to zero; (ii) whenever a chromosome block in $\mathbf{R}^{\circ}\mathbf{R}$ was too large (more than 2000 markers), markers within the chromosome were first pruned out in a sliding-window approach (1000 markers by window, overlap of 500 markers), by solving $(\mathbf{R}^{\circ}\mathbf{R})\mathbf{w} = 1$ restricted to each window separately and discarding markers with weights equal to zero.

Training, Testing, and Validation

We considered three types of sets, for training and testing, regarding observations at maternal parents. The HS-family BLUPs used for training and/or testing the prediction model could be either from Arlington, Marshfield, or the average over the two locations

Prediction procedures were evaluated using prediction accuracy estimated in five-fold cross-validation, replicated 20 times. Given a random partition of instances in five subsets of similar size, four subsets were used for training and the remaining subset was used for testing. For each of the five subsets used sequentially for testing, prediction accuracy was computed as the Pearson coefficient of correlation between "observed" and predicted HS-family BLUPs.

Finally, expected gains from GEBV-based selection were computed in a manner parallel to those for field-based selection systems. Expected gain from GEBV-based selection on an individual-plant basis was computed using the formula given by [11] for ΔG_{INDGS} and one of the two different selection intensities (0.10 or 0.01). Both computations were expressed as a percentage of the HSF selection method applied to sward-plot biomass yield (direct selection) with a selection intensity of 0.10. Phenotypic selection using the HSF method was based on 5 years per cycle (20 years for four generations or recombination events). Genomic selection was based on a 5-year training cycle followed by three selection cycles of 1 year each (October to October), resulting in a total of 8 years for four generations or recombination events.

Results

Biomass Yield and Field Traits

There was significant genetic variability for biomass yield within this switchgrass population, with significant family × method and family × location interactions, as indicated by relatively narrow confidence intervals (Table 1). The family × year interactions were the least important, largely because repeated harvests on the same plots are correlated with each other (Table 2). Harvesting sward plots for 2 years is probably sufficient to obtain the most efficient biomass yield estimates, because the additional information provided in the third and fourth years does not offset the added cost from extending cycle time by another year or two.

This was also confirmed by an analysis of the correlation coefficients between BLUP values from individual years, locations, and plot methods (Table 3). The highest mean and maximum correlation coefficients were observed between individual years within locations and methods. Correlation coefficients, both the maximum and the mean, were substantially lower for location-to-location comparisons within plot methods. Finally, the lowest values, by far, were the correlation coefficients between plot methods (spaced plants, row plots, and sward plots). Figure 1 provides a visualization of the family × location interactions for sward plots (Fig. 1a) and row plots (Fig. 1b), clearly showing a modest positive relationship between the two Wisconsin locations and no relationship between Wisconsin and Nebraska. For BLUP values computed over years and locations of the three plot methods, there were small positive relationships of row-plot biomass and spaced-plant biomass with sward-plot biomass yield (Fig. 2). The low correlations between plot methods indicate that either row plots or spaced plants will require an additional

Table 1 Estimates of random effects associated with half-sib families of WS4U-C2 switchgrass evaluated across multiple plot methods (M), locations (L), and years (Y), including 95% confidence intervals (CI)

Source of variation	df	Estimate	Lower 95% CI	Upper 95% CI
Half-sib family (F)	143	0.1053	0.0269	6.6166
F × M	429	0.3259	0.1542	1.0867
F × L/M	286	0.4010	0.2230	0.9242
F × Y/M	715	0.0617	0.0067	$> 10^{11}$
F × (L × Y)/M	572	0.5922	0.2841	1.9044
Residual	4410	13.2694	12.7172	13.8585

advantage, such as increased selection intensity or reduced cycle time, to be advantageous.

Narrow-sense heritability for biomass was the lowest for sward plots and row plots compared to spaced plants, but this may be due to the fact that spaced plants were only established at one location (Table 4). The largest family × location interaction was observed for row plots, which were established at the two most divergent locations: Arlington, WI and Mead, NE, two distinctly different hardiness zones and climatic zones. Nevertheless, the family variance was significant for all three plot types, indicating that any of these methods can be used effectively in a switchgrass breeding program. However, the strong family × method interaction (Tables 1 and 3) clearly indicates that these three measures of biomass should probably be considered as potentially different traits, i.e., that simply considering them as one trait measured on three different plot types may be too simplistic of an interpretation.

Heading and anthesis dates were considered as possible indirect selection criteria in this study, because of the important role they have played in switchgrass breeding programs during the past 20 years [44]. Genetic variation and heritability estimates were high for both heading date scored on row plots and anthesis date scored on spaced plants (Table 5). Heading date was far less susceptible to family × location interactions than biomass yield, with family × year interactions as the most important source of environmental interaction for these traits.

Table 2 Correlation coefficients of best linear unbiased predictors (BLUPs) for biomass yield of switchgrass sward plots over 4 years with BLUPs based on 1, 2, or 3 years of biomass yield measurements from two locations

Location	First-year BLUP	Two-year BLUP	Three-year BLUP
4-year Arlington BLUP	0.462	0.847	0.894
4-year Marshfield BLUP	0.654	0.814	0.870
4-year BLUPs over two locations	0.578	0.851	0.899

Table 3 Minimum (Min), maximum (Max), and mean correlation coefficients between best linear unbiased predictors (BLUPs) for biomass yield of switchgrass measured under 14 combinations of plot method, location, and year (n = number of individual correlation coefficients pooled together)

Correlation coefficient	Number	Min	Max	Mean	SE
Between years (within locations and methods)	15	− 0.222	0.748	0.224	0.063
Between locations (within methods)	24	− 0.259	0.575	0.134	0.036
Between plot types (within the Arlington location)	20	− 0.178	0.184	0.028	0.024

Expected gains for various indirect selection criteria are presented in Table 6. Despite the advantages offered by using

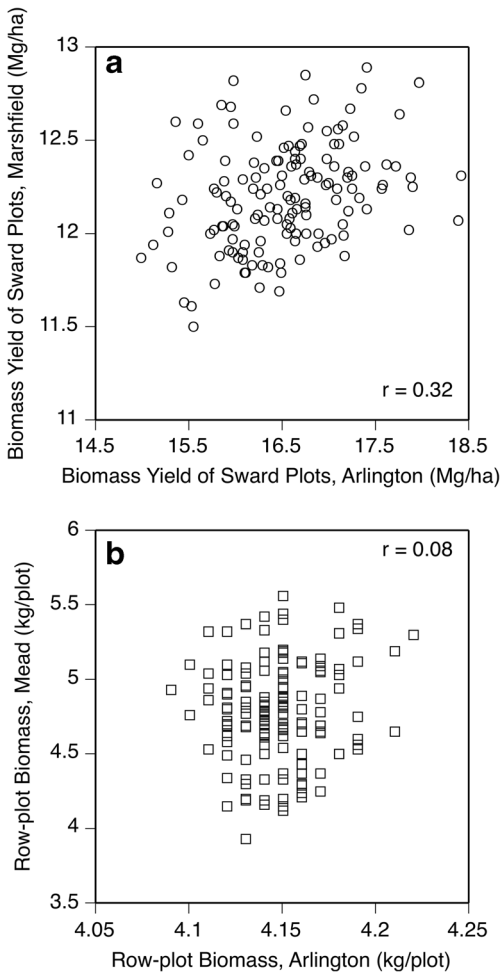

Fig. 1 Scatterplots and linear regressions of best linear unbiased predictors (BLUPs) for biomass yield of 144 switchgrass half-sib families measured at two locations in either sward plots or row plots. **a** Sward BLUPs from Marshfield vs. sward BLUPs from Arlington. **b** Row-plot BLUPs from Mead, NE vs. row-plot BLUPs from Arlington, WI

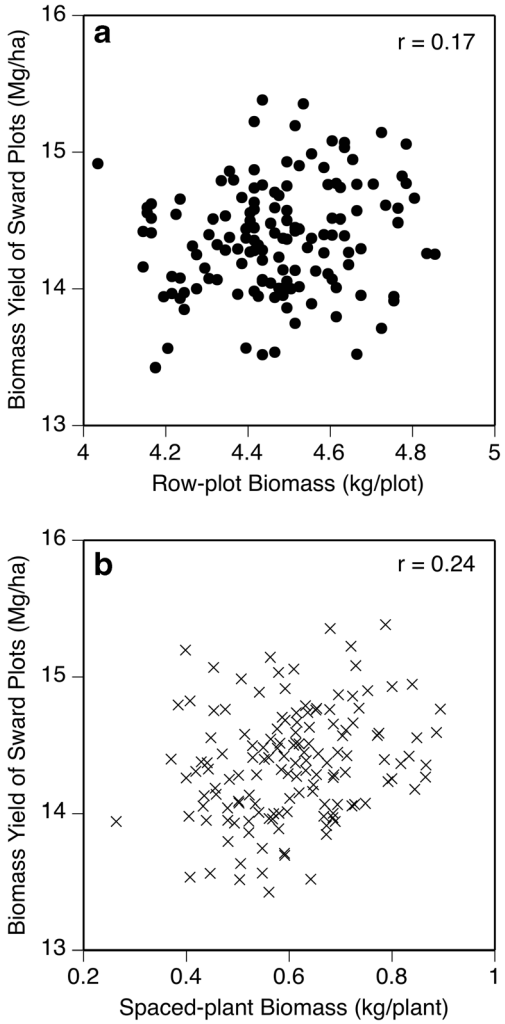

Fig. 2 Scatterplots and linear regressions of best linear unbiased predictors (BLUPs) for biomass yield of 144 switchgrass half-sib families measured in sward plots, row plots, or spaced plants. **a** Sward BLUPs based on two locations and 4 years vs. row-plot BLUPs based on two locations and 2 years. **b** Sward BLUPs based on two locations and 4 years vs. row-plot BLUPs based on one locations and 2 years

row plots or spaced plants (increased selection intensity and cycle time reduced from 4 to 2 years), indirect selection was less efficient than HSF selection on a family mean basis, even with a 10 times increase in selection intensity. Conversely, when two selection trials are established concomitantly, one to measure biomass yield of sward plots and one to measure biomass, heading date, or anthesis date of individual plants, as was done by [7], all of these options demonstrated expected gains greater than HSF for biomass yield selection alone, with spaced-plant biomass and row-plot heading date giving the greatest expected gains.

Genomic Estimated Breeding Values

Two genomic DNA platforms were investigated: ECS and GBS. The ECS platform was based on a defined gene space of approximately 60 MB of the genome with relatively little missing data due to the use of specific probe sequences [18]. Conversely, the GBS platform utilized DNA from the entire genome, but contained massive amounts of missing data or missing cells in the data matrix. There was a clear difference in performance between the two DNA platforms, with the ECS platform giving consistently higher prediction accuracies for GEBVs compared to GBS (Table 7). Several marker-data transformation and imputation methods were applied in an attempt to improve the accuracy of GEBV prediction from GBS, all without any success, even though, among GBS-based procedures, relatively good results were obtained with HMM imputation and *Cor* marker-data transformation, as well as with MI imputation and *LD* marker-data transformation. Likewise, several prediction models were applied to the phenotypic and genomic data without any substantial benefit compared to the classical GBLUP method. Lastly, even though there was significant family × location interaction for biomass yield of sward plots, genomic predictions could be made across locations without any substantial loss in accuracy (Table 8).

The expected gain from selection based on GEBV was 184% relative to the expected gains from HSF selection, i.e., almost double the gain expected from the field-based program. If funds are sufficient to conduct DNA analysis of a larger population, potentially allowing an increase in selection intensity of 10 times, expected gains for GEBV-based selection increase to 278%, almost a 3 times improvement over field-based HSF selection. While the accuracies of GEBV prediction in this population are low, the gains result from the drastic reduction in average cycle time: one 3-year training period followed by three 1-year selection cycles. However, this selection scheme has one potential caveat: decay of linkage disequilibrium over generations will erode the prediction accuracy and decrease the realized gains made in the second and third selection cycles [13, 33].

Discussion

Plant breeders are always looking for ways to increase the efficiency of breeding programs. If we loosely define efficiency as rate of gain per unit of time, this gives us two mechanisms for increasing efficiency: increasing the rate of gain per se or shortening the cycle time, or perhaps both. Breeding perennial energy grasses officially began in 1992 [45], but is really a tangential application of forage breeding, an activity that traces its roots to Wales and Denmark in the late nineteenth century [46]. Much like breeding forage, fodder, and pasture crops, a very large proportion of breeding activity in the energy grasses takes place in spaced-planted nurseries. The clear advantage of this method is that it allows breeders and other staff members to make repeated observations and

Table 4 Estimates of random effects associated with half-sib families of WS4U-C2 switchgrass evaluated in three plot types with multiple blocks (B), locations (L), and years (Y), including lower and upper 95% confidence limits

Source of variation	Sward-plot biomass yield			Row-plot biomass			Spaced-plant biomass		
	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Family (F)	0.995	0.576	2.116	0.071	0.034	0.238	0.022	0.016	0.033
F × L	0.000	–	–	0.035	0.008	6.685	NA		
F × B/L	2.786	2.051	4.005	0.501	0.404	0.636	0.018	0.014	0.024
F × Y	0.031	0.024	$> 10^{20}$	0.000	–	–	0.004	0.003	0.009
F × L × Y	1.585	0.849	3.941	0.035	0.011	0.518	NA		
Residual	21.309	19.955	22.807	1.058	0.962	1.170	0.019	0.016	0.022
Heritability	0.33			0.25			0.66		

NA not applicable

measurements on individual plants over the course of several years without fear of contamination. Even seed set on fertile plants prior to biomass harvests can be controlled with tillage or pre-emergence herbicides. Every breeder uses spaced-planted nurseries for one purpose or another. The critical question here is whether or not they are being overused or misused.

Spaced plantings consist of a highly unnatural environment, in which individual genotypes are not allowed to compete with each other, taking away an extremely essential component a sward environment or community established in a production field. Indeed, the distinction between environments with vs. without interplant competition can be used as a method of classifying traits as simple or complex traits. Simple traits are those that can be effectively measured on spaced plants and genetic gains are realized in sward plots [47]. Numerous examples, far too many to cite here, include flowering time, simple morphological traits such as stem and leaf characteristics, biomass quality traits (e.g., lignin, ash, or N concentrations), pest resistances, and some stress tolerances [47]. Conversely, for traits such as biomass yield, the literature contains numerous examples of failures to make breeding progress when selection is conducted on spaced plants and evaluations are conducted in sward plots to simulate realistic production conditions [46–48]. Biomass yield should be classified as a complex trait, i.e., a trait which cannot be reliably measured on spaced plantings without interplant competition.

This conclusion, derived from numerous observations on a wide range of species [e.g.,49–52], is supported by the results of the present study. The correlations between sward-plot biomass yield, row-plot biomass, and spaced-plant biomass were all positive and low, indicating that each of these three traits can be used to make progress toward improving sward-plot biomass yields. This helps to explain the fact that much of the early progress in breeding for increased biomass yields of switchgrass for biomass production were accomplished in spaced-plant or row-plot nurseries [11, 44]. With a small positive correlation between these three plot types, genetic progress is possible, but the results in Table 6 clearly show that gains resulting from selection without interplant competition will be lower than those using sward plots as a direct measure of biomass yield. Even allowing for the fact that these indirect selection measures involve a halving in cycle or generation time and a possible 10 times increase in selection intensity, selection for increased biomass or later flowering on spaced plantings or row plots is less efficient than direct selection for biomass yield of sward plots!

Table 5 Estimates of random effects associated with heading and anthesis dates for half-sib families of WS4U-C2 switchgrass evaluated in two plot types with multiple blocks (B), locations (L), and years (Y), including lower and upper 95% confidence limits

Source of variation	Row-plot heading date			Spaced-plant anthesis date		
	Estimate	Lower	Upper	Estimate	Lower	Upper
Family (F)	10.071	7.709	13.718	7.934	5.911	11.212
F × L	0.207	0.032	$> 10^{5}$	NA		
F × B/L	6.270	5.382	7.400	0.000		
F × Y	2.634	2.006	3.614	3.509	2.560	5.110
F × L × Y	0.000			NA		
Residual	6.876	6.270	7.574	4.477	3.984	5.067
Heritability	0.75			0.76		

NA not applicable

Table 6 Predicted correlated responses to selection for indirect selection criteria using two selection intensities, all expressed as a percentage of the expected response to direct selection for sward-plot biomass yield with a selection intensity of 10%

Selection method and traits[a]	Selection intensity	Indirect selection criterion (trait X)			
		Row-plot biomass	Spaced-plant biomass	Row-plot heading date	Spaced-plant anthesis date
	---------- % ----------				
Individual selection for trait X	0.10	27	63	56	25
Individual selection for trait X	0.01	40	95	85	38
Combined selection for traits X and Y	0.10	110	124	121	109
Combined selection for traits X and Y	0.01	115	136	132	114

[a] Individual selection is based on selection of individual plants within either spaced plantings or row plots and requires 2 years per cycle, 1 year for establishment and 1 year for selection and recombination in situ. Combined selection (AWFX-HS of Casler and Brummer, 2008) is based on measuring biomass yield on sward plots and trait X on either row plots or spaced plants of the same families established at the same time: 1 year for establishment, 2 years for data collection, and a fourth year for creation of a recombination block by transplanting selected individual, followed by seed production

Given the positive correlations for biomass production among the three plot types, there is a distinct advantage to a combined selection protocol that utilized sward plots to identify the best families and spaced plants or row plots to identify the best plants within those families (Table 6). Utilization of the vast amounts of additive genetic variance

Table 7 Mean prediction accuracy for genomic predicted breeding values, across marker-data types, marker-data transformations, and prediction models for sward plot yield in WI averaged over years and locations

Data type	Data transformation[a,b,c,d]	Prediction model				(Mean)
		GBLUP	RKHS	BayesB	RF	
ECS	Base	0.235	0.234	0.214	0.237	0.230
	PCA	0.235	0.234	0.185	0.027	0.170
	Cor	0.207	0.196	0.177	0.188	0.192
	LD	0.197	0.209	0.183	0.177	0.192
	(Mean)	0.219	0.218	0.190	0.157	0.098
GBS_HMM	Base	0.145	0.117	0.103	− 0.001	0.091
	PCA	0.145	0.117	0.086	0.040	0.097
	Cor	0.159	0.105	0.144	0.133	0.135
	LD	0.062	0.079	0.045	0.031	0.054
	(Mean)	0.128	0.105	0.095	0.051	0.098
GBS_MI	Base	0.109	0.067	0.105	0.183	0.116
	PCA	0.109	0.067	0.087	0.117	0.095
	Cor	− 0.023	0.098	− 0.020	0.137	0.048
	LD	0.168	0.152	0.167	0.208	0.174
	(Mean)	0.091	0.096	0.085	0.161	0.098
GBS_RFI	Base	0.131	0.051	0.081	0.082	0.086
	PCA	0.131	0.051	0.073	0.075	0.083
	Cor	0.101	0.061	0.015	0.069	0.062
	LD	0.114	0.050	0.102	0.122	0.097
	(Mean)	0.119	0.053	0.068	0.087	0.098

GBLUP genomic best linear unbiased predictor, *RKHS* reproducing kernel Hilbert space, *BayesB* Bayesian B, *RF* random forest, *ECS* exome capture sequencing, *GBS* genotype-by-sequencing, *HMM* hidden Markov model, *MI* mean imputation, *RFI* random forest imputation

[a] Base = standard input data based on allelic dosages

[b] PCA = input data are principal components of the data matrix

[c] Cor = input data are marker variables scaled through a correlation matrix

[d] LD = input data are marker variables adjusted for redundancy due to linkage disequilibrium

Table 8 Mean prediction accuracy for genomic predicted breeding values, across training schemes for sward plot yield measured in Arlington and/or Marshfield averaged over years, using the optimal prediction procedure identified in Table 7: *Base*—GBLUP with exome capture sequencing

Training location	Testing location		
	Arlington	Marshfield	Combined
Arlington	0.200	0.235	0.229
Marshfield	0.206	0.209	0.219
Combined	0.214	0.231	0.235

within half-sib families is critically ignored in the HSF breeding system [47, 48, 53, 54]. Simple traits such as plant biomass or flowering time can be used to apply positive and meaningful selection pressure within families, but at a fairly high cost in either labor or time. This combined selection method, AWFX-HS [15], involves the use of simultaneous and concomitant experiments—a sward-plot study to measure biomass yield for 2 years and a spaced-plant or row-plot nursery from which to select the best plants within the best families after the yield evaluation has been completed. The intensive labor requirement arises from the need to score or measure all plants on all families prior to the completion of the sward-plot yield trial, a massive amount of investment into data collection on plants and families that have no chance of being selected for recombination. The alternative is to delay data collection on the spaced-plant or row-plot nursery until after the highest-yielding families have been identified, but this would increase cycle or generation time by at least 1, possibly 2 years. The disadvantage of this would be to decrease the relative efficiencies in Table 6. For example, just adding 1 year to the cycle time would decrease the values on the bottom line of Table 6 by 20% each, resulting in values of 92, 109, 105, and 91%, respectively. Essentially, a 1-year delay would defeat the whole purpose of this selection scheme.

Strictly from a historical standpoint, we already know that "Godzilla" selection (PS) as described above [7], selection for later flowering time [11], and selection for spaced-plant biomass [2] are all effective for increasing biomass yield of sward plots. However, the demonstrated inefficiency of these methods compared to direct selection for biomass yield in this population brings us full circle to the dilemma of the best way forward. It is clear that genomic methods offer a clear advantage to improve the efficiency of breeding for increased biomass yield in three ways: decreasing average cycle time, increasing selection intensity, and allowing meaningful selection pressure to be exerted on all of the additive genetic variance within a population, i.e., both among and within families [12, 13]. As a result, genomic methods allowed for possibly doubling or tripling the expected gains compared to field-based HSF selection, despite accuracies of GEBV for sward plot yield being quite low in the WS4U-C2 population. Such accuracies could not be improved substantially by alternate marker-data transformations or prediction models, compared to a standard GBLUP approach. In our analysis, the ECS platform yielded markedly higher prediction accuracies compared to the less expensive GBS platform. These results should not generalize to all contexts, but they do suggest that investing in more expensive genotyping platforms may be worthwhile regarding prediction accuracy and genetic gains. Selection methods based on DNA markers could be further improved using more sophisticated approaches, such as selection methods that combine both genomic data and phenotypic data together into selection indices [12]. Furthermore, the relatively low prediction accuracies observed for the WS4U-C2 population, while still of a magnitude to make genomic selection advantageous, are clearly lower than those observed for switchgrass populations that contain larger amounts of genetic variability [13].

The two protocols for determining genotype used in this study illustrated very distinct sequencing approaches. The two methods differed by the type of representation reduction and were characterized by very different sequencing depths. The GBS protocol was designed as a cost-efficient option for genotyping and was multiplexed at 96 times, whereas ECS was multiplexed at 12 times. As a result of these major differences in multiplexing and genome space reduction, more markers were discovered by ECS, compared to GBS, resulting in better genome coverage (average physical distance between markers were 9.1 and 100.8 kb for ECS and GBS, respectively). Given the short extent of LD in WS4U-C2 [12], the larger marker density in ECS certainly contributed to the higher genomic prediction accuracies in comparison to GBS. Naturally, the superiority of ECS over GBS may also have been due to the higher uncertainty in genotype calling and the higher proportions of missing values in GBS [25, 55, 56].

As a result of the low sequencing depth in GBS, a liberal threshold on proportions of missing values was used in GBS, so as to retain a large number of markers. In GBS, thresholds for maximum proportions of missing values at 60, 40, and 20% resulted in only 3652; 1039 and 100 markers selected, respectively, while a threshold at 80% resulted in 10,856 markers selected. Besides, thresholds lower than 80% did not result in significant gains in prediction accuracy, based on GBS_MI and GBLUP: prediction accuracies were 0.109, 0.061, -0.033, and 0.122 for thresholds at 80, 60, 40, and 20% missing values, respectively. Notably, the gain in prediction accuracy realized for < 20% missing values was non-significant ($p = 0.50$ based on a paired t test). Therefore, we chose to use 80% as a cutoff for missing values, in order to maximize the amount of marker information available for imputation and model calibration. However, deeper

analyses on this aspect of genomic prediction protocols may prove useful, as optimization of such cutoffs by imputation approaches, prediction models, and/or marker-data transformations could probably yield further increases in prediction accuracy.

Conclusions

In conclusion, breeders of perennial energy grasses are strongly encouraged to cease or drastically reduce the use of spaced plantings with no interplant competition or row plantings with minimal interplant competition for the purpose of selection for increased biomass yield. We furthermore recommend that biomass measurements from sward plots and biomass or non-competitive plots be treated as different traits, e.g., biomass yield on a per-hectare basis can be extrapolated from sward plots, but not from plants within no or minimal interplant competition. Instead, we suggest the term "plant biomass" or something similar to distinguish this as a different trait than what is measured on sward plots created by either drill planting or broadcast seeding.

We also recommend that breeders of perennial energy grasses strongly consider modifications to improve breeding efficiency by one or more of the following three factors: (1) decrease cycle time by spending only the amount of time necessary to obtain accurate assessments of phenotype, (2) increase selection intensity by evaluating larger populations for shorter periods of time and with simpler and more efficient phenotypic evaluation methods, and (3) devise selection schemes that capture ALL of the additive genetic variance in the population undergoing selection. Genomic selection may not be an option for every energy grass breeder, but as costs continue to decrease and technology continues to become more mainstream, more breeding programs are going to have such access and be able to fund this type of activity.

Acknowledgements This research was funded in part by the Agriculture and Food Research Initiative Competitive Grant No. 2011-68005-30411 from the USDA National Institute of Food and Agriculture (CenUSA); by the U.S. Department of Energy, BER Office of Science, Great Lakes Bioenergy Research Center Grant DE-FC02-07ER64494; and by congressionally allocated funds through USDA-ARS. We thank Nick Baker and Joe Halinar (USDA-ARS, U.S. Dairy Forage Center), Michael Bertram (Arlington Agricultural Research Station, University of Wisconsin), and Jason Cavadini (Marshfield Agricultural Research Station, University of Wisconsin) for research support of field operations. We thank Dr. Kenneth Vogel (retired) and Steven Masterson (USDA-ARS, Lincoln, NE) for research support of field operations at Mead, NE. The authors thank the University of Wisconsin Biotechnology Center Bioinformatics facility for providing genotype-by-sequencing analysis services.

References

1. Perrin RK, Vogel KP, Schmer MR, Mitchell RB (2008) Farm-scale production cost of switchgrass for biomass. Bio Energy Res 1:91–97
2. Casler MD, Vogel KP (2014) Selection for biomass yield in upland, lowland, and hybrid switchgrass. Crop Sci 54:626–636
3. Missaoui AM, Fasoula VA, Bouton JH (2005) The effect of low plant density on response to selection for biomass production in switchgrass. Euphytica 142:1–12
4. Rose LW IV, Das MK, Fuentes RG, Taliaferro CM (2007) Effects of high- vs low-yield environments on selection for increased biomass yield of switchgrass. Euphytica 156:407–415
5. Henderson CL (1984) Applications of linear models in animal breeding. University of Guelph, Guelph, Ontario, Canada
6. Price DL, Casler MD (2014a) Inheritance of secondary morphological traits for among-and-within-family selection in upland tetraploid switchgrass. Crop Sci 54:646–653
7. Casler MD (2010) Changes in mean and genetic variance during two cycles of within-family selection in switchgrass. Bio Energy Res. 3:47–54
8. Bhandari HS, Saha MC, Fasoula VA, Bouton JH (2011) Estimation of genetic parameters for biomass yield in lowland switchgrass (*Panicum virgatum* L.) Crop Sci 51:1525–1533
9. Das MK, Fuentes RG, Taliaferro CM (2004) Genetic variability and trait relationships in switchgrass. Crop Sci 44:443–448
10. Boe AR, Beck DL (2008) Yield components of biomass in switchgrass. Crop Sci 48:1306–1311
11. Price DL, Casler MD (2014b) Divergent selection for secondary traits in upland tetraploid switchgrass and effects on sward biomass yield. BioEnergy Res 7:329–337
12. Resende RMS, Casler MD, de Resende MV (2014) Genomic selection in forage breeding: accuracy and methods. Crop Sci 54:143–156
13. Ramstein GP, Evans J, Kaeppler SM, Mitchell RB, Vogel KP, Buell CR, Casler MD (2016) Accuracy of genomic prediction in switchgrass (*Panicum virgatum* L.) improved by accounting for linkage disequilibrium. Genes, Genomes, Genet 6:1049–1062
14. Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. Statistical Science 6:15-51
15. Casler MD, Brummer EC (2008) Theoretical expected genetic gains for among-and-within-family selection methods in perennial forage crops. Crop Sci 48:890–902
16. Falconer DS (1989) Introduction to quantitative genetics. Longman Scientific & Technical, Essex, England. 3rd ed
17. Evans J, Crisovan E, Barry K, Daum C, Jenkins J, Kunde- 921 Ramamoorthy G, Nandety A, Ngan CY, Vaillancourt B, Wei CL 922 (2015) Diversity and population structure of northern switchgrass as 923 revealed through exome capture sequencing. Plant J 84(4):800–815
18. Evans J, Kim J, Childs KL, Vaillancourt B, Crisovan E, Nandety A, Gerhardt DK, Richmond TA, Jeddeloh JA, Kaeppler SM, Casler MD, Buell CR (2014) Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass Panicum virgatum. Plant J. https://doi.org/10.1111/tpj.12601
19. Poland J, Brown PJ, Sorrells ME, Jannink J-L (2012a) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7(2):e32253
20. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25

21. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079
22. Martin ER, Kinnamon D, Schmidt MA, Powell E, Zuchner S, Morris R (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26(22): 2803–2810
23. Jiang H, Lei R, Ding S-W, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15(1):1
24. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635
25. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2):e90346
26. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359
27. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink J-L (2012b) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome J 5(3):103. https://doi.org/10.3835/plantgenome2012.06.0006
28. Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3):250–255
29. Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. Am J Hum Genet 98(1):116–126
30. Stekhoven DJ, Bühlmann P (2012) MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118
31. Li G, Serba DD, Saha MC, Bouton JH, Lanzatella CL, Tobias CM (2014) Genetic linkage mapping and transmission ratio distortion in a three-generation four-founder population of Panicum virgatum (L.) G3: Genes| Genomes| Genet 4(5):913–923
32. Pudovkin A, Zaykin D, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics 144(1):383–387
33. Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397. https://doi.org/10.1534/genetics.107.081190
34. Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res 91(01):47–60
35. Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829
36. Gianola D, van Kaam JB (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178(4):2289–2303
37. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
38. Pérez P, de los Campos G (2014) Genome-wide regression & prediction with the BGLR statistical package. Genetics 114:164442
39. Shen X, Alam M, Fikse F, Rönnegård L (2013) A novel generalized ridge regression method for quantitative genetics. Genetics 193(4): 1255–1268
40. Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2):215–223
41. Gu C, Ma P (2005) Optimal smoothing in nonparametric mixed-effect models. Ann Stat 33:1357–1379
42. Liaw A, Wiener M (2002) Classification and regression by randomForest. R news 2(3):18–22
43. Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91(6):1011–1021. https://doi.org/10.1016/j.ajhg.2012.10.010
44. Casler MD (2012) Switchgrass breeding, genetics, and genomics. In: Monti A (ed) Switchgrass. Springer, New York, pp 29–54
45. Sanderson MA, Adler PR, Boateng AA, Casler MD, Sarath G (2007) Switchgrass as a biofuels feedstock in the USA. Can J Plant Sci 86:1315–1325
46. Casler MD, Pedersen JF, Eizenga GC, Stratton SD (1996) Germplasm and cultivar development. In: Moser LE et al (eds) Cool-season forage grasses. American Society of Agronomy, Madison, pp 413–469
47. Casler MD, van Santen E (2010) Breeding objectives in forages. In: Boller B et al (eds) Handbook of plant breeding, vol 5. Fodder crops and amenity grasses. Springer, NY, pp 115–136
48. Wilkins PW, Humphreys MO (2003) Progress in breeding perennial forage grasses for temperate agriculture. J Agric Sci Camb 140: 129–150
49. Hayward MD, Vivero JL (1984) Selection for yield in *Lolium perenne*: II. Perform spaced plant sel under compet conditions Euphytica 33:787–800
50. Carpenter JA, Casler MD (1990) Divergent phenotypic selection response in smooth bromegrass for forage yield and nutritive value. Crop Sci 30:17–22
51. Annicchiarico P (2006) Prediction of indirect selection for seed and forage yield of lucerne based on evaluation under spaced planting. Plant Breed 125:641–643
52. Waldron BL, Robins JG, Peel MD, Jensen KB (2008) Predicted efficiency of spaced-plant selection to indirectly improve tall fescue sward yield and quality. Crop Sci 48:443–449
53. Humphreys MO (2005) Genetic improvement of forage crops—past, present, and future. J. Agric. Sci. Camb. 143:441–448
54. Vogel KP, Pedersen JF (1993) Breeding systems for cross-pollinated perennial grasses. Plant Breed Rev 11:251–274
55. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9(1):e1003215. https://doi.org/10.1371/journal.pgen.1003215
56. Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics 193:1073–1081