Subsampling and DNA pooling can increase gains through genomic selection in switchgrass

Neal Wepking Tilhou1 | Michael D. Casler2

1 Department of Agronomy, University of Wisconsin, 1575 Linden Dr, Madison, WI 53706, USA
2 U.S. Dairy Forage Research Center, USDA–ARS, 1925 Linden Dr, Madison, WI 53706-1108, USA

Correspondence
Neal W. Tilhou, Department of Agronomy, University of Wisconsin, 1575 Linden Dr, Madison, WI, 53706, USA.
Email: wepking@wisc.edu

Assigned to Associate Editor David Hyten.

Abstract
Genomic selection (GS) can accelerate breeding cycles in perennial crops such as the bioenergy grass switchgrass (Panicum virgatum L.). The sequencing costs of GS can be reduced by pooling DNA samples in the training population (TP), only sequencing TP phenotypic outliers, or pooling candidate population (CP) samples. These strategies were simulated for two traits (spring vigor and anthesis date) in three breeding populations. Sequencing only the outlier 50% of the TP phenotype distribution resulted in a penalty of <5% of the predictive ability, measured using cross-validation. Predictive ability also decreased when sequencing progressively fewer TP DNA pools, but TPs constructed from only two phenotypically contrasting DNA samples retained a mean of >80% predictive ability relative to individual TP sequencing. Novel group testing methods allowed greater than one CP individual to be screened per sequenced DNA sample but resulted in a predictive ability penalty. To determine the impact of reduced sequencing, genetic gain was calculated for seven GS scenarios with variable sequencing budgets. Reduced TP sequencing and most CP pooling methods were superior to individual sequence-based GS when sequencing resources were restricted (2,000 DNA samples per 5-yr cycle). Only one scenario was superior to individual sequencing when sequencing budgets were large (8,000 DNA samples per 5-yr cycle). This study highlights multiple routes for reduced sequencing costs in GS.

1 | INTRODUCTION

Switchgrass (Panicum virgatum L.) is a perennial bioenergy grass that is native to the United States and is capable of producing large quantities of low-input biomass (Casler, 2012; Casler et al., 2018). Bioenergy crops have an important role in ameliorating climate change because they are storable energy sources and are feedstocks for carbon-negative technologies (Gelfand et al., 2020; Sanchez et al., 2018). The economic viability of switchgrass can be improved with higher yielding cultivars (Perrin et al., 2008, 2017). However, yield improvement in switchgrass is slow because a selection cycle requires multiple years for field evaluation and one year for recombination of selected individuals. Currently, one field selection cycle requires approximately 5 years (Casler & Ramstein, 2018; Price & Casler, 2014). Selection using yield surrogates, such as flowering time or winter survival, can shorten the...
breeding cycle to 2–3 years (Casler, 2019; Poudel et al., 2019, 2020). Adoption of genomic selection (GS) may instead allow one selection cycle per year (Fiedler et al., 2018; Lipka et al., 2014; Ramstein et al., 2016).

Genomic selection allows a breeder to isolate high breeding value individuals using genetic markers instead of field evaluations (Crosa et al., 2017; Jannink et al., 2010; Meuwissen et al., 2001). Genomic selection is valuable for breeding programs where field evaluations are expensive, time-intensive, or incapable of allowing within-family selection (Heffner et al., 2010; Resende et al., 2014). During GS, a breeder assembles a training population (TP) with marker and field performance data and also obtains marker data from a population of candidate genotypes (the candidate population; CP).

To select elite individuals, marker-trait relationships from the TP are applied to the CP to calculate genomically estimated breeding values (GEBVs). Selection cycles using GS require genotyping of hundreds or thousands of DNA samples and this expense may limit the utility of GS for many breeding programs. Critically, this includes many small or low-resource breeding programs (Delannay et al., 2012; Ribaut & Ragot, 2019; Ribaut et al., 2010; Coe et al., 2020). Strategies that reduce sequencing expenses for GS could expand GS adoption and improve breeding progress.

Pooled DNA sequencing (PDS), that is, pooling DNA samples from the TP into groups of similar trait values, reduces costs with moderate reductions in predictive ability (Alexandre et al., 2019; Baller et al., 2020; Bell et al., 2017; Henshall et al., 2012; Reverter et al., 2016). Sequencing expenses can also be reduced by only including individuals with outlier trait values in the TP. This has previously been referred to as selective genotyping but will be referred to as outlier sequencing (OS) in the current study (Lander & Botstein, 1989; Li et al., 2019; Morton & Howarth, 2005). In yellow croaker (Larimichthys polyactis Bleeker), predictions that omitted the central 50% of the TP (50% OS) only lost 4% accuracy relative to sequencing the complete TP (Dong et al., 2016). Reduced sequencing strategies such as OS and PDS can be used together and could reduce TP sequencing, data management, and sample processing costs.

Breeders must also obtain GEBVs from large CPs during a GS cycle. Greater cost savings can occur through CP PDS since the CP often contain a greater number of individuals than TPs for breeding programs using forward selection. Minimizing sequencing expenses in the CP can be viewed as a group testing problem (Dorffman, 1943; Hughes-Oliver, 2006; McMahan et al., 2012). Specifically, group testing attempts to pool samples with the goal of minimizing the number of tests (DNA samples sequenced) per individual assessed while maintaining adequate accuracy. Two simple strategies for pooled testing involve one- and two-dimensional CP PDS. One-dimensional PDS involves randomly assigning individuals to pools of equal sizes. This results in a potential savings of $1-(1/n)$, where $n$ is the number of individuals per CP pool. This reduces sequencing expenses but confounds GEBVs of individuals sharing a DNA pool. To remedy this, adaptive testing (AT) can be carried out, where a second round of individual sequencing resolves individual GEBVs within DNA pools with promising GEBVs. Two-dimensional PDS allows a breeder to obtain individual GEBVs from a single sequencing cycle (Bilder & Tebbs, 2012; McMahan et al., 2012; Westreich et al., 2008). Two-dimensional PDS involves assigning individuals to a square array where PDS occurs for individuals sharing a row or a column. This allows a potential savings of $1-(2n/n^2)$ where $n$ is the number of individuals in each row or column of the array. Group testing methods have not been applied to GS CPs and may provide a beneficial balance between sequencing inputs and prediction accuracy.

To evaluate reduced sequencing GS strategies in switchgrass, this study simulates a range of reduced sequencing strategies across three switchgrass breeding groups that were evaluated for two traits: anthesis date and spring vigor. The simulations determined the utility of both PDS and OS in the TP with variable marker read depth, varying outlier proportions of TP individuals sequenced, and a variable number of TP DNA pools. Simulations of the CP varied the number of CP individuals per pool, varied the AT proportions, and compared one-dimensional and two-dimensional PDS. This study also compared the potential genetic gain of seven reduced sequencing methods relative to individual sequencing GS for two switchgrass breeding populations.

### Core Ideas

- Pooling DNA and subsampling can reduce genomic selection sequencing expenses.
- Group testing can reduce sequencing requirements when screening new genotypes.
- Reduced sequencing can improve genetic gain in a simulated switchgrass breeding program.

## 2 MATERIALS AND METHODS

### 2.1 Field data collection

This experiment used unreplicated, randomized spaced plants as the experimental unit. Individuals were established from greenhouse-grown seedlings that were planted in Arlington, WI (43°18’ N; 89°21’ W), in rows with 90 cm between rows and 30 cm within rows with 25 plants per row. To simplify data collection, alternating seedlings were killed in spring 2018, resulting in 60 cm spacing between individuals and up...
to 13 plants per row. The nurseries were blocked into groups of 20 rows from the same breeding group. A subset of individuals was selected for sequencing, resulting in blocks consisting of between 80 and 128 unreplicated individuals. Anthesis date was collected for 3 yr (2018–2020) and spring vigor was recorded during spring 2019 and 2020. Spring vigor was recorded on a scale from 0 to 20, with 20 indicating no visible damage and 0 indicating mortality. Fall vigor, scored from 0 to 5, was also collected as a covariate for the subsequent spring vigor scores. Anthesis date was recorded as the date in which at least 50% of panicles contained 50% protruding anthers. Measurements were pruned if they were determined to be biologically implausible. Specifically, individual anthesis dates that differed by greater than 2 wk from the anthesis dates recorded during the two other years were removed. Similarly, individuals with spring vigor scores differing by >15 between the 2 yr were removed.

2.2 | Germplasm

The breeding groups are upland individuals (n = 529), lowland individuals (n = 302), and a stabilized upland–lowland hybrid population (n = 422). All individuals are allotetraploids that have strongly disomic inheritance (Okada et al., 2010). The breeding groups are undergoing selection for cold tolerance and late flowering. Upland individuals are the result of one cycle of selection for late flowering and tall plants from the WS4U population (Casler et al., 2006). Lowland individuals have undergone one or two cycles of selection for winter survival and are derived from either ‘Kanlow’ or are of mixed ancestry from the southern Great Plains. Population structure exists within the lowland population because it contains individuals with no history of random mating with the Kanlow-derived selections (Supplemental Figure S3). The hybrid nursery individuals are from the ‘Liberty’ hybrid cultivar derived from upland (‘Summer’) and lowland (Kanlow) parents (Vogel et al., 2014).

2.3 | Generation of best linear unbiased predictors

Best linear unbiased predictors (BLUPs) based on field performance data were used to assess genomic prediction accuracy for each population-trait combination. Various models were tested for each trait to determine the most informative model based on the Akaike information criterion. The following linear model was used to calculate BLUPs of anthesis date for each genotype:

\[ y_{jkl} = \mu + p_j + M_k + b_l + \epsilon_{jkl} \]

where \( y_{jkl} \) is the observed flowering time measurement, \( p_j, M_k, \) and \( b_l \) are the effects of an individual (random), year (fixed), and block (fixed) respectively, and \( \epsilon_{jkl} \) are residuals (i.i.d. ∼ N(0,\( \sigma^2_e \))). Genotypic effect variance–covariance structures were estimated using marker relationships (\( p_j \sim \text{MVN}(0, A\sigma^2_g) \)). The A matrix of genomic relationships between individuals was calculated using A.mat in the rrBLUP package (Endelman, 2011).

The following linear model was used to calculate BLUPs of spring vigor for each genotype:

\[ y_{jkl} = \mu + p_j + M_k + b_l + V_l + \epsilon_{jkl} \]

where \( y_{jkl} \) is the observed spring vigor score, \( p_j, M_k, \) and \( b_l \) are the effects of an individual (random), year (fixed), block (fixed), and fall vigor score (fixed), respectively, and \( \epsilon_{jkl} \) are residuals. The model assumed unequal variances between sampling years and fall vigor score categories. Genotypic effect variance–covariance structures were estimated using markers (\( p_j \sim \text{MVN}(0, A\sigma^2_g) \)). The A matrix of genomic relationships between individuals was calculated using the A.mat function in the rrBLUP package (Endelman, 2011). Estimates of BLUP values, variances, and repeatability were obtained using the R package ‘sommer’ (Covarrubias-Pazaran, 2016). Repeatability was calculated as follows:

\[ \rho = (\sigma^2_G + \sigma^2_{PE})/(\sigma^2_G + \sigma^2_{PE} + \sigma^2_e/n) \]

where \( \sigma^2_G \) represents the genotypic variance and \( \sigma^2_{PE} \) represents the permanent environmental variance. In this experimental design, genotypic and permanent environmental variances are confounded. The symbol \( \sigma^2_e \) represents the error variance and \( n \) is the number of measurements per individual. Genotypic variance was calculated as the sum of additive genotypic and residual genotypic variance using the linear models presented above with an additional genotype effect with a diagonal covariance matrix to estimate residual genotypic variance.

2.4 | DNA collection, sequencing, and SNP calling

Tissue samples were collected and freeze dried for DNA extraction in September 2018. Genotyping-by-sequencing occurred on an Illumina sequencer through the University of Wisconsin Biotechnology Center. Raw GBS sequences were processed at the Wisconsin–Madison’s Bioinformatics Resource Center using a standard TASSLE-GBS pipeline described as in Glaubitz et al. (2014). Briefly, the bar-coded sequence read outputs were collapsed into a set of unique sequence tags with counts. Tags were aligned to the reference genome (P. virgatum v5.1; Lovell et al., 2021),
assigning each tag to a position with the best unique alignment. The occupancies of tags for each sample were observed from barcode data. Resulting files were used to call single-nucleotide polymorphism (SNP) markers at the tag locations on the genome. Resulting SNPs were filtered for minimum taxa coverage (0.01), minimum locus coverage (0.2), and minor allele frequency (0.05). Imputation of missing markers (37%) was accomplished using Beagle (Ayres et al., 2012). The final dataset consisted of 530,810 SNPs.

2.5 | Genomic prediction methods

All genomic prediction models were constructed using ridge-regression BLUP (rrBLUP), specifically, using the following model:

\[ y = W \mathbf{g} + \mathbf{e} \]

where \( y \) is a vector of the observed BLUPs of individuals or the mean BLUPs of individuals within a pool. The matrix \( W \) is the design matrix relating genotypes to observations, \( \mathbf{G} \) is the biallelic marker matrix coded as \(-1, 0, 1\) for individual marker dosage or a frequency (0 to 1) for pool marker dosage. The vector \( \mathbf{u} \) contains marker effects and \( \mathbf{e} \) is a vector of residuals (Endelman, 2011). Marker-based predictive ability was calculated as the Pearson correlation coefficient between the GEBVs of a randomly omitted CP and their BLUP values.

2.6 | Pooled DNA simulations

Each simulation broadly followed the standard structure of (a) omitting a subset for cross-validation, (b) assembling the TP to generate the marker–trait relationships, and (c) assessing the predictive ability within the CP. All simulations used cross-validation with 15% omitted from the potential TP and each possible combination of the below variables were replicated 50 times. Three sets of simulations were run evaluating OS and PDS within TPs and one set of simulations was run evaluating PDS strategies in CPs. Simulations were carried out on five trait–population combinations. This includes anthesis date in the upland, lowland, and hybrid breeding groups and winter survival in the hybrid and lowland breeding groups. The upland population has strong winter survival and was not evaluated for spring vigor.

The first TP simulation set compared the predictive ability of a complete TP and a 50% OS TP using individual genotypes for both the TP and CP. These individual sequence simulations were also replicated with different TP sizes based on randomly selecting a subset of the available individuals prior to TP construction.

The second set of TP simulations determined the impact of PDS pool size and sequencing effort, which was measured as the number of markers and depth of sequencing. The TPs were evenly divided into either four, eight, or 50 DNA pools based on similar BLUP values for PDS. The number of markers varied by randomly selecting 0.1, 0.5, 1, 5, 25, 50, or 100% of the available markers for analysis. Variable read depth was simulated by randomly sampling from the called alleles of individuals selected for the simulated DNA pool.

The third cycle of TP simulations varied both TP PDS degree and TP OS percentage. The TP was first limited to outlier individuals of between 10 and 100%, by increments of 10%, of the potential TP. Specifically, OS preferentially includes outlier individuals and therefore omitted individuals from the center of the distribution first. The remaining outlier individuals were then pooled equally into two, six, 20, 50, or 100 DNA samples based on similar BLUP values. Simulations of OS and PDS that resulted in fewer than two individuals per DNA pool were omitted.

All simulations of CP PDS used marker–trait relationships from a TP constructed from the entire BLUP distribution pooled into 50 DNA samples based on similar BLUP values. The CP simulations evaluated the effect of pool size, AT proportion, and the sample pooling strategy (one-dimensional and two-dimensional). The CP PDS varied from two to six individuals per pool. One-dimensional PDS randomly assigned individuals into equally sized DNA pools. A GEBV was then generated from the mean marker frequencies of all individuals within each pool and the GEBV was assigned to each individual within the pool. Adaptive testing was carried out on one-dimensional PDS scenarios and involved a second round of sequencing of individuals within the top 20, 40, 60, or 80% GEBVs of the initial DNA pools. Adaptive testing requires an additional \( n-1 \) sequences for each pool since the GEBV of the final individual can be determined by comparing the other \( n-1 \) individual GEBVs in the pool with the pool-wide GEBV.

For two-dimensional PDS, samples were assigned to one location in a square array of DNA samples with the number of rows and number of columns equal to the pool size. Arrays were constructed from random individuals in the CP until no further complete arrays can be assembled. The remaining individuals were discarded for that simulation. The simulated DNA samples consisted of pools containing individuals sharing a row or column in each array. Marker frequencies for each sample in a marker array were then estimated in a two-step process. First, any individual in a row or column that is uniform (homozygous) for a marker was assigned the marker value. Second, the remaining marker estimates are calculated as the mean of the residual read values in the row and column of the sample (Supplemental Figure S1). Individual marker values estimated from the arrays were used to generate GEBV for individuals based on TP marker effects.
TABLE 1  Breeding scenarios with variable reduced sequencing strategies applied to the training population or the candidate population that will be compared with genomic selection with individually sequenced training and candidate populations

| Breeding scenario | Training population | Candidate population |
|-------------------|---------------------|----------------------|
| 50OS              | 50% outlier individuals sequenced | Individually sequenced candidate population |
| PDS50             | 50 pooled DNA sequences | Individually sequenced candidate population |
| 1D2N              | 50 pooled DNA sequences | One-dimensional pooled DNA sequencing with two individuals per pool |
| 1D2N40AT          | 50 pooled DNA sequences | One-dimensional pooled DNA sequences with two individuals per pool and 40% secondary adaptive testing |
| 1D3N60AT          | 50 pooled DNA sequences | One-dimensional pooled DNA sequences with three individuals per pool and 20% secondary adaptive testing |
| 2D3N              | 50 pooled DNA sequences | Two-dimensional pooled DNA sequencing with three individuals per pool |
| 2D4N              | 50 pooled DNA sequences | Two-dimensional pooled DNA sequencing with four individuals per pool |

2.7 Genetic gain estimates

Progress was calculated based on predictive abilities from simulations of two breeding population–trait combinations: anthesis date selection in the upland switchgrass breeding population and spring vigor in the hybrid switchgrass breeding population. Seven breeding scenarios were compared with GS with individually sequenced TP and CP (Table 1). Scenario 50OS and PDS50 evaluated TP OS and PDS, respectively. The remaining scenarios used 50 PDS TPs and varied CP reduced sequencing methods. Each breeding scenario attempted to maximize genetic gain using a set number of sequenced DNA samples, either 2,000; 4,000; or 8,000 sequenced DNA samples per 5-yr breeding cycle. Each GS breeding cycle assumed a 2-yr evaluation of the TP and three recurrent selection years of GS. Sequencing resources were first allocated to construct the TP. After, the remaining sequencing resources were evenly among the CPs of the 3 yr of recurrent GS.

To describe a complex example scenario, the 2,000 DNA sample Scenario 1D2N40AT first uses 50 pooled DNA sequences during TP construction then allocates the remaining 1,950 DNA samples evenly to the three CP cycles. These 650 DNA sequences are then used for either the initial round of pooled sequencing or for the 40% AT. Effectively, the initial round of pooled sequencing costs 0.5 DNA sequence per evaluated individual. Deconvoluting the top 40% of pools costs an additional 0.5 DNA sequence per high GEBV pool. Therefore, the overall cost is 0.7 \((0.5 + (0.4 \times 0.5))\) sequenced sample per evaluated individual and 928 individuals can be evaluated per CP cycle for Scenario 1D2N40AT.

Genetic gain was calculated based on the breeders equation (Lush, 1943). The true additive variance is equal within each population–trait combination. Therefore, the two major variables impacting genetic gain are selection accuracy and selection intensity. Each breeding scenario assumed 100 selected individuals per GS year. Therefore, the savings from either PDS or OS allowed reallocation of sequencing effort to increase CP size. Greater CP size results in greater selection intensity. Each scenario was reported as a proportion of gain relative to individually sequenced GS. The experiment assumed a 25% loss in predictive ability per generation without an updated TP (Neyhart et al., 2017).

3 RESULTS

All breeding group and trait combinations had moderate to high repeatability (>0.50; Table 2). Unexpectedly, spring vigor repeatability in the hybrid group (0.55) was lower than GS predictive ability (0.76). Omitting the central 50% of the TP (50% OS) did not alter the predictive ability of spring vigor in the lowland population but reduced hybrid predictive ability by 3.9% (0.76–0.73; Table 2). A similar minor decline occurred in anthesis predictions, where predictive ability was reduced by 3.1% (0.63–0.61), 1.6% (0.60–0.59), and 11.4% (0.61–0.54) for upland, lowland, and hybrid breeding groups as a result of OS 50%. The predictive ability loss resulting from OS 50% was not exacerbated by decreasing TP size (Figure 1). In simulations including only 20% of the potential TP \((n = 51, n = 72, n = 90\) for lowland, upland, and hybrid, respectively), using 50% OS TP reduced anthesis date predictive ability (Figure 1). At 20% TP size, hybrid spring vigor predictive ability increased from 0.34 to 0.39 with 50% OS. At 20% TP size, lowland and upland breeding group predictive ability did not change as a result of OS 50% for anthesis date prediction (0.54 and 0.47, respectively). In spring vigor predictions using TP 20%, 50% OS reduced the hybrid group predictive ability by 4.8% (0.62–0.59) and the lowland group predictive ability by 2.0% (0.49–0.48).

Pooling DNA from genotypes into 50 DNA samples (50 PDS TP) with similar trait values prior to sequencing tended to have reduced predictive ability relative 50% OS (Table 2). Anthesis date predictive ability changed by −18.3% (0.61–0.50) and −3.3% (0.63–0.61) for hybrid and upland groups, respectively. No reduction occurred as a result of 50 PDS TP.
TABLE 2  The sample size, repeatability, and genomic predictive ability for five breeding group–trait combinations with standard error (repeatability) or standard deviations (predictive ability, \( n = 100 \)). Repeatability is the ratio between genotypic variance (additive and nonadditive) and phenotypic variance (genotypic and error variance).

| Breeding group | Trait       | \( n \) | Repeatability | Individual sample predictive ability | 50 PDS\(^a\) TP\(^b\) predictive ability | 50% OS\(^c\) TP predictive ability |
|----------------|-------------|---------|---------------|--------------------------------------|----------------------------------------|----------------------------------|
| Upland         | Anthesis date | 529     | 0.77 ± 0.02   | 0.63 ± 0.07                          | 0.60 ± 0.06                           | 0.61 ± 0.06                      |
| Lowland        | Anthesis date | 300     | 0.65 ± 0.04   | 0.60 ± 0.09                          | 0.60 ± 0.11                           | 0.59 ± 0.11                      |
| Lowland        | Spring vigor | 302     | 0.73 ± 0.12   | 0.62 ± 0.10                          | 0.59 ± 0.10                           | 0.62 ± 0.08                      |
| Hybrid         | Anthesis date | 422     | 0.78 ± 0.02   | 0.61 ± 0.08                          | 0.50 ± 0.12                           | 0.54 ± 0.10                      |
| Hybrid         | Spring vigor | 422     | 0.55 ± 0.09   | 0.76 ± 0.05                          | 0.70 ± 0.06                           | 0.73 ± 0.06                      |

\(^a\)PDS, pooled DNA sequencing.  
\(^b\)TP, training population.  
\(^c\)OS, outlier sequencing.

FIGURE 1  Mean predictive ability of genomic prediction within three breeding populations (upland, \( n = 529 \); hybrid, \( n = 422 \), and lowland, \( n = 302 \)) and two traits (spring vigor and anthesis date). Simulations used 50 or 100% of the available breeding population using outlier sequencing (OS). Simulations were replicated with subsets of 20, 40, 60, 80, and 100% of the available breeding group. Predictive ability is mean. Each bar represents 50 simulations and error bars represent one standard deviation from the mean.

for lowland anthesis date prediction (0.60). Spring vigor predictive ability was reduced in the hybrid group by 7.9% (0.76–0.70) and the lowland group by 4.8% (0.62–0.59).

Simulations that reduced depth or marker number of pooled DNA samples found similar predictive ability trends between simulations with a small number of DNA pools (containing hundreds of individuals per pool) and greater numbers of DNA pools (containing fewer individuals per pool; Supplemental Figure S2). Broadly, differential predictive abilities resulting from pooling occurred if <5% of the initial markers were included. This difference was the most apparent in upland anthesis predictions, which indicated a reduction in predictive ability when the simulated read depth per marker was below five reads and the TP consisted of either four or eight DNA pools.

Simulations that varied TP PDS and OS rates found that TPs could maintain comparable predictive abilities relative to individual sequencing (>90%) with either >50 DNA pools sequenced or >50% OS (Figure 2). Variable rates of degradation occurred in different trait-population combinations below those thresholds but even extreme forms of OS and PDS resulted in >70% of individually sequenced TP predictive ability (Figure 2). Relative to individual TP sequencing, the most extreme degree of PDS (2 DNA Pool TPs) simulations of anthesis date genomic prediction resulted in a predictive ability decrease of 26.2% (0.61–0.45), 23.3% (0.60–0.54), and 6.3% (0.63–0.59) for hybrid, lowland, and upland breeding groups.
The Plant Genome

**FIGURE 3** Mean predictive ability of genomic prediction simulations applied to a candidate population (CP) with contrasting DNA pooling strategies. Genotypes evaluated per sequenced DNA sample ($p$) used in the CP is related to pool size ($n$) where $p = n$ for one dimensional pools and $p = n^2/2n$ for two dimensional pools. Error bars represent one standard deviation from the mean. Points represent the mean value of 100 simulations and error bars represent one standard deviation from the mean. Horizontal dashed bars represent the mean predictive ability using individual training population and CP sequencing for anthesis (0.63, 0.61, and 0.60 for upland, hybrid, and lowland, respectively) and spring vigor (0.76 and 0.62 for hybrid and lowland, respectively).

Groups, respectively. Simulations of spring vigor 2 PDS TP decreased predictive ability by 14.4% (0.76–0.65) and 22.6% (0.62–0.48) for hybrid and lowland breeding groups, respectively. Similarly, the most extreme form of OS (10% OS) simulations of anthesis date prediction resulted in a predictive ability decrease of 31.1% (0.61–0.42; 6 PDS), 11.7% (0.60–0.53; 6 PDS), and 17.4% (0.63–0.52; 20 PDS) for hybrid, lowland, and upland breeding groups, respectively. The simulations of spring vigor genomic prediction using OS 10% decreased predictive ability by 21.0% (0.76–0.60; PDS 6) and 9.7% (0.62–0.56; PDS 6) for hybrid and lowland breeding groups, respectively.

One- and two-dimensional CP PDS resulted in reductions in predictive ability with increasing number of pooled genotypes per sequenced DNA samples (Figure 3). Where one- and two-dimensional pooling resulted in equivalent sequencing costs, predictive abilities were comparable for both traits in lowland and upland breeding groups (Figure 3). However, two-dimensional pooling resulted in reduced predictive ability for the hybrid breeding group. This effect was greatest for hybrid population spring vigor predictions (Figure 3). Incorporating AT into one-dimensional CP PDS allowed less than two genotypes to be evaluated per sequenced DNA sample, which resulted in predictive ability approaching to individual sequencing (Figure 4). For example, one-dimensional PDS with two individuals per DNA pool and AT >20% attained, on average, >95% predictive ability relative to individual CP sequences while requiring between 0.7 and 0.9 sequences per CP individual.

Multiple reduced sequencing GS scenarios were compared with GS using individual sequencing. Scenarios implementing TP OS and PDS (Scenario 50OS and Scenario PDS50) resulted in greater genetic gain with the upland anthesis date, but resulted in similar or reduced genetic gain in Liberty spring vigor (2000 DNA samples per cycle scenarios; Figure 5). Reallocating sequencing resources from the TP to the CP resulted in an increase in selection intensity that more than compensated for the minor reduction in predictive ability observed in these simulations. With greater sequencing resources (4,000 DNA samples and 8,000 DNA samples), Scenario 50OS and Scenario PDS50 were inferior to individually sequenced TP GS. The majority of the CP PDS strategies resulted in superior gains relative to individually sequenced GS when sequencing is limited to 2,000 DNA samples. For these scenarios, the loss in predictive ability resulting from CP PDS was also offset by the benefit of a larger CP. One exception was the poor performance of hybrid two-dimensional PDS (Scenarios 2D3N and 2D4N).

Relative genetic gain of CP PDS scenarios decreased in the 4,000 and 8,000 DNA sequence scenarios (Figure 5). However, Scenario 1D2N40AT (Table 1; Figure 5) was superior to individual sequencing GS in all scenarios. Scenario 1D2N40AT resulted in 14.9, 6.5, and 2.7% greater gains for hybrid spring vigor and 21.9, 9.8, and 5.7% greater for anthesis date (for 2,000; 4,000; and 8,000 DNA sequence budgets, respectively). If Scenario 1D2N40AT is recalculated to only use the DNA sequencing resources required to match the
genetic gain of an individually sequenced CP, it results in a sequencing savings of 35.0, 23.9, and 15.3% for hybrid spring vigor and 43.4, 33.6, and 21.8% for anthesis date (for 2,000; 4,000; and 8,000 DNA sequence budgets, respectively). Genetic gain from two-dimensional CP PDS was consistently inferior to one-dimensional CP PDS and inferior to individual sequenced GS for all but the 2,000 DNA sequence upland anthesis breeding scenario.

4 | DISCUSSION

4.1 | Reducing sequencing in training populations

These simulations highlight an economic tradeoff between predictive ability and sequencing costs when constructing a TP. Improved sequencing technology may lessen this tradeoff through continued reductions in per-base-pair sequencing costs (Bhat et al., 2016; Muir et al., 2016). However, GS performance plateaus at low or moderate marker densities (Abed et al., 2018; Fiedler et al., 2018; Werner et al., 2018; Zhang et al., 2019) and per-sample expenses currently account for the majority of genotype-by-sequencing costs for a GS switchgrass sample. These per-sample costs are primarily labor and reagents which may not decline in cost going forward. Therefore, strategies which reduce per-sample sequencing expenses may be valuable for breeding programs.

These switchgrass GS simulations found that outlier individuals contain the greatest information value for GS and individuals with moderate trait values provide progressively less value (Lander & Botstein, 1989; Li et al., 2019; Morton & Howarth, 2005). This is intuitive since the tails of the trait distribution contain the greatest share of variance in a population. In this assessment, switchgrass individuals in the central 50% of the distribution contribute <5% of the predictive ability on average, a trend that remains true even in TPs containing <100 individuals (Figure 1). Therefore, a proportion of moderate-trait-value TP individuals may be economically omitted from sequencing for many breeding programs using GS. Even in the case of multitrait breeding programs, strategies could be devised to isolate individuals providing the least information. For example, Alexandre et al. (2019) pooled individuals using Z scores, which normalized each trait of interest. This strategy could be extended to also exclude individuals with the lowest absolute mean Z scores.

The use of PDS for TP construction uncouples the linear relationship between TP size and sequencing expenses. This strategy is most valuable in systems where diversity can be affordably generated. Since PDS samples cluster a heterogeneous population into a smaller number of DNA pools sharing a trait value, phenotypic information is lost, as fewer PDS
samples are included in the TP (Figure 2). The granular trait information lost as a result of TP PDS resulted in a moderate to severe predictive ability penalty, but even extreme applications (10% OS TP or 2 PDS TP) resulted in, at worst, ∼70% of the maximum predictive ability (Figure 2).

The use of PDS to affordably extract trait–marker relationships from potentially large TPs can enable new methods of TP construction. As is described in Reverter et al. (2016), sampling genotypes from commercial settings can facilitate data collection from the true production environment rather than from experiments that are proxies of commercial settings. While commercial sampling limits the experimental methods available to a breeder, production populations sizes are usually much larger than what could be evaluated in an experiment. Provided measurement errors are uncorrelated with genotypic values, errors in marker frequency within these large pools will regress to zero with increasing sample size. Since DNA pooling can extract marker information from a TP with sequencing costs unrelated to the number of individuals in a TP, screening very large commercial populations may be a viable method for TP construction or supplementation of the experimental TP. This strategy is limited to easily phenotyped traits but could be useful for challenging traits such as drought tolerance, resistance to a disease outbreak, or long-term persistence.

The most obvious drawback to these reduced sequencing strategies for TPs is that the DNA pools are only valuable for the current trait of interest. There are also statistical limitations imposed by OS and PDS as used in the current study. Both PDS and OS can limit the detection of population structure, can prevent accurate estimation of genetic variances, and can prevent accurate scaling of the GEBVs. Although population structure cannot be detected or accounted for using the usual methods in PDS TPs, population structure can be recognized through inflated predictive abilities (Berro et al., 2019; Fiedler et al., 2018). If population structure in a breeding population is associated with the trait of interest, artificially high predictive abilities can be observed simply by assigning individuals to correct subpopulations. This can be seen in the lowland breeding group, which is known to include multiple unmixed subpopulations. Since one or more of the lowland switchgrass subpopulations vary for the traits measured, unrealistically strong predictive abilities are observed even with 99% of the markers omitted from analysis (Supplemental Figure S1).

There are many facets to breeding programs that these TP simulations did not evaluate. These include breeding strategies such as pedigrees, multitrait GS, dominance or epistatic variance modeling, and genotype-by-environment effects. In addition, these simulations only used DNA pools with uniform numbers of individuals in the TP (and CP, below). The ability to incorporate PDS samples with variable number of individuals into a TP could improve the flexibility of these methods. Lastly, switchgrass is an outcrossing species, and the breeding populations used in this experiment were created by recurrent selection and random recombination (Casler, 2012). Therefore, the assumption of random mating is generally valid for these breeding programs. Breeding programs that do match the assumption of random mating may have more spurious relationships when TP sequencing is reduced through PDS or OS.

### 4.2 Pooled candidate populations

The likelihood that an untested genotype is elite is low in plant breeding. Therefore, efficient field experiments maximize the number of untested genotypes evaluated while minimizing the early investment in each genotype (Moehring et al., 2014; Williams et al., 2011). Similar strategies have not been applied to resource allocation within GS CPs. Incorporating group testing methods may reduce the sequencing expenses per CP individual. However, CP PDS resulted in a steep trade-off between predictive ability and savings in the current simulations (Figures 3 and 4).

Many group testing strategies assume skewed sample distributions with a low proportion of outliers (Chen & Hwang, 2008). Therefore, the normal distribution of GEBVs makes multidimensional group testing ineffective for accurately identifying high GEBV individuals. Instead of decoding two-dimensional arrays using the GEBV of pooled DNA samples, two-dimensional pooled samples in the current experiment were decoded at the scale of genetic markers prior to GEBV calculation since rare markers can be effectively identified within a two-dimensional array (Supplemental Figure S1). Reasonable predictive abilities were obtained using this strategy in two of the three breeding groups, but the hybrid breeding group two-dimensional PDS predictions performed poorly (Figure 3). The reason for this is unknown but could be related to the population’s upland–lowland ancestry (Ramstein et al., 2016).

The trade-off between predictive ability and sequencing effort was improved by the use of AT in one-dimensional pools. The one-dimensional AT simulations used the initial PDS sequencing run to isolate pools with promising candidates and used AT to identify the specific elite individuals. The majority of the sequencing savings were provided through PDS, but scenarios with initial PDS of three and four individuals per pool resulted in subtly lower plateaus of predictive ability after AT (Figure 4). This likely is due to the failure to identify pools containing elite individuals during the initial PDS stage.

These simulations applied simple proof-of-concept strategies for CP DNA pooling and decoding. More advanced methods of pool construction and array decoding are possible (Balding et al., 1996; Chen & Hwang, 2008; McMahan...
et al., 2017). For example, individuals could also be replicated across the one- or multidimensional pools. Replication would increase complexity but can reduce the risk of losing promising genotypes during the initial cycle of PDS. These simulations also did not assess AT in two-dimensional pools. Sequencing an individual within a two-dimensional pool array could improve the marker decoding accuracy across the entire two-dimensional pooling array. Lastly, pool decoding accuracy could be improved using population genetic information such as linkage disequilibrium or phased parental sequence information (Technow & Gerke, 2017). Since individual sequencing of large CPs are costly investments, and most CP individuals are rejected, strategies that further improve PDS CP methods could result in substantial cost reductions for GS.

4.3 Limited sequencing breeding scenarios

The reduced sequencing strategies used for TP construction were beneficial when sequencing was limited to 2,000 DNA samples (Figure 5; Scenarios 500S and PDS50). Reallocating sequencing savings from the TP to CP has the greatest benefit with small CPs. Under conditions with greater sequencing resources, the reallocation of TP sequencing resources became a penalty in genetic gain. Although not simulated in the current study, both OS and PDS will likely have the greatest benefit when the methods allow TPs to contain more individuals while maintaining constant sequencing expenditures.

BROADLY, the predictive ability penalties of CP PDS resulted in poor outcomes when the proportion of CP individuals evaluated per DNA sequence became greater than two (Figure 4). By using one-dimensional CP PDS with two individuals per pool and adaptive testing on 40% of the highest GEBV pools, scenario 1D2N40AT sequenced 0.7 DNA samples per individual in the CP (or 1.43 individuals per DNA sample). After including reallocation from TP PDS, the final predictive ability penalty was only 0.01 (1.6%) while evaluating a 52% larger CP than if individual sequencing was used. This trade-off was beneficial even in high-resource scenarios. Since the 50 PDS TP resulted in a predictive ability penalty under these conditions (Scenario PDS50), greater benefits may have been observed if Scenario 1D2N40AT used individual TP DNA sequencing.

The methods of the current study attempted to evaluate the trade-off between sequencing expenses and predictive ability without assigning program-specific costs to sequencing or field evaluation, both of which can vary substantially between breeding programs. It also omitted other potential expenses in the breeding process. For example, these budgets assumed that new genotypes in the CP could be affordably produced. This fact is unlikely to be the case in systems where inbreeding or manual pollination are used. However, in switchgrass, the production of new CP genotypes is currently affordable.

These simulated genetic gain calculations also did not include the impact of predictive ability variance in the simulations. The impact of minor predictive ability fluctuations could have major impacts on genetic gain. Despite these limitations, these results provide promising evidence that group testing methods can to reduce GS CP sequencing costs and improve genetic gain.

5 CONCLUSIONS

Reduced sequencing TP strategies, such as OS and PDS, can allow breeders to balance the effort, accuracy, and expenses of phenotyping and genotyping when constructing a TP. Preferentially sequencing TP outlier individuals can reduce sequencing expenses while maintaining acceptable predictive ability even with small initial TP sizes. Pooling TP individuals based on similar trait values can also reduce sequencing costs and can be useful when the population available for the TP is too large to individually sequence. Group testing methods provided a novel route for reduced CP sequencing costs. Broadly, multiple CP pooling strategies were evaluated but resulted in similar trade-offs between sequencing savings and predictive ability. This study used simple CP pooling methods and there are multiple routes for improvement during DNA pool construction and decoding. To determine the benefits of reduced sequencing methods, genetic gain was compared among seven scenarios using variable reduced sequencing methods. Reduced TP sequencing methods produced greater genetic gain relative to individually sequenced TPs only in low-sequencing conditions (2,000 DNA samples per cycle). Reduced sequencing strategies for both TP and CP were less beneficial when greater sequencing resources were assumed (4,000 DNA and 8,000 DNA samples per cycle). However, one reduced sequencing breeding scenario used CP pooling with adaptive testing to use only 0.7 DNA sequences per individual in the candidate population. This scenario reallocated the savings to expand the size of the CP and resulted in 5–22% greater genetic gain relative to individual GS sequencing.

DATA AVAILABILITY STATEMENT

A marker matrix and trait BLUPs will be posted to Dryad Digital Repository.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Neal Tilhou: Conceptualization, data collection and curation, formal analysis, writing. Michael Casler: Project conception and direction, funding acquisition, supervision, methodology, review of manuscript.
ACKNOWLEDGMENTS

The authors thank Hari Poudel Ph.D., Julian Cooper, John Raasch, Joseph Halinar, and Emily Twohig for their help with field maintenance and data collection, and Professor Joshua Tebbs (Department of Statistics, University of South Carolina) and Professor Christopher McMahan (Department of Mathematical Sciences, Clemson University) for statistical advice. Also, I would like to thank Dean Sanders, Ph.D., for his advice regarding bioinformatics and sequencing. This material is based upon work supported in part by the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018409. This research was also supported by congressionally allocated funds to the USDA–ARS, U.S. Dairy Forage Research Center.

ORCID

Neal Wepking Tilhou https://orcid.org/0000-0002-5221-1809

REFERENCES

Abed, A., Pérez-Rodríguez, P., Crossa, J., & Belzile, F. (2018). When less can be better: How can we make genomic selection more cost-effective and accurate in barley? *Theoretical and Applied Genetics*, 131, 1873–1890. https://doi.org/10.1007/s00122-018-3120-8

Alexandre, P. A., Porto-Neto, L. R., Karaman, E., Lehnhert, S. A., & Reverter, A. (2019). Pooled genotyping strategies for the rapid construction of genomic reference populations. *Journal of Animal Science*, 97, 4761–4769. https://doi.org/10.1093/jas/skz344

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Hulsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., & Suchard, M. A. (2012). BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology*, 61, 170–173. https://doi.org/10.1093/sysbio/syr100

Balding, D. J., Bruno, W. J., Torney, D. C., & Knill, E. (1996). A comparative survey of non-adaptive pooling designs. In T. Speed & M. S. Waterman (Eds.), *Genetic mapping and DNA sequencing* (pp. 133–154). Springer.

Ballaré, J. L., Kuchman, S. D., Kuehn, L. A., & Spangler, M. L. (2020). Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. *Journal of Animal Science*, 98, skaa184. https://doi.org/10.1093/jas/skaa184

Bell, A. M., Henshall, J. M., Porto-Neto, L. R., Dominik, S., McCulloch, R., Kijas, J., & Lehnhert, S. A. (2017). Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. *Genetics, Selection, Evolution*, 49, 28. https://doi.org/10.1186/s12711-017-0303-8

Berro, I., Lado, B., Nalin, R. S., Quincke, M., & Gutiérrez, L. (2019). Training population optimization for genomic selection. *The Plant Genome*, 12, 190028. https://doi.org/10.3835/plantgenome2019.04.0028

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushqaq, M., Jain, N., Singh, P. K., Singh, G. P., & Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics*, 7, 221. https://doi.org/10.3389/fgene.2016.00221

Bilder, C. R., & Tebbs, J. M. (2012). Pooled testing procedures for screening high volume clinical specimens in heterogeneous populations. *Statistics in Medicine*, 31, 3261–3268. https://doi.org/10.1002/sim.5334

Casler, M. D. (2012). Switchgrass breeding, genetics, and genomics. In A. Monti (Ed.), *Switchgrass* (pp. 29–53). Springer.

Casler, M. D. (2019). Selection for flowering time as a mechanism to increase biomass yield of upland switchgrass. *BioEnergy Research*, 13, 100–108. https://doi.org/10.1007/s12155-019-10044-3

Casler, M. D., & Ramstein, G. P. (2018). Breeding for biomass yield in switchgrass using surrogate measures of yield. *BioEnergy Research*, 11, 1–12. https://doi.org/10.1007/s12155-017-9867-y

Casler, M., Vogel, K., & Beal, A. (2006). Registration of WS4U and WS8U switchgrass germplasms. *Crop Science*, 46, 998–999. https://doi.org/10.2135/cropsci2005.04-0020

Casler, M. D., Vogel, K. P., Lee, D. K., Mitchell, R. B., Adler, P. R., Sulc, R. M., Johnson, K. D., Kallenbach, R. L., Boe, A. R., Mathison, R. D., Cassidy, K. A., Min, D. H., Crawford, J., & Moore, K. J. (2018). 30 years of progress toward increased biomass yield of switchgrass and big bluestem. *Crop Science*, 58, 1242–1254. https://doi.org/10.2135/cropsici2017.12.0729

Chen, H. B., & Hwang, F. K. (2008). A survey on nonadaptive group testing algorithms through the angle of decoding. *Journal of Combinatorial Optimization*, 15, 49–59. https://doi.org/10.1007/s10878-007-9083-3

Coe, M. T., Evans, K. M., Gasic, K., & Main, D. (2008). Plant breeding capacity in U.S. public institutions. *Crop Science*, 60, 2373–2385. https://doi.org/10.1002/csc2.20227

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS ONE*, 11, e0156744. https://doi.org/10.1371/journal.pone.0156744

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Vashney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, 22, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Delannay, X., Mclaren, G., & Ribaut, J. M. (2012). Fostering molecular breeding in developing countries. *Molecular Breeding*, 29, 857–873. https://doi.org/10.1007/s11032-011-9611-9

Dong, L., Xiao, S., Chen, J., Wan, L., & Wang, Z. (2016). Genomic selection using extreme phenotypes and pre-selection of SNPs in large yellow croaker (*Larimichthys crocea*). *Marine Biotechnology*, 18, 575–583. https://doi.org/10.1007/s10126-016-9718-4

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14, 436–440. https://doi.org/10.1214/aoms/1177731363

Endelman, J. B. (2011). Ridge regression and other kernels for genomic prediction capacity in U.S. public institutions. *The Plant Genome*, 12, 2373–2385. https://doi.org/10.1007/s00122-011-0173-8

Fiedler, J. D., Lanzatella, C., Edmé, S. J., Palmer, N. A., Sarath, G., Quincke, M., Fideli, F., & Tobias, C. M. (2018). Genetic mapping and DNA sequencing funds to the USDA–ARS, U.S. Dairy Forage Research Center.
Gelfand, I., Hamilton, S. K., Kravchenko, A. N., Jackson, R. D., The- len, K. D., & Robertson, G. P. (2020). Empirical evidence for the potential climate benefits of decarbonizing light vehicle transport in the U.S. with bioenergy from purpose-grown biomass with and without BECCS. *Environmental Science & Technology*, 54, 2961–29742. https://doi.org/10.1021/acs.est.9b07019

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & BUCKLER, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLOS ONE*, 9, e90346. https://doi.org/10.1371/journal.pone.0090346

Heffner, E. L., Lorenz, A. J., Jannink, J. L., & Sorrells, M. E. (2010). Plant breeding with genomic selection: Gain per unit time and cost. *Crop Science*, 50, 1681–1690. https://doi.org/10.2135/cropsci2009.11.0662

Henshall, J. M., Hawken, R. J., Dominik, S., & Barendse, W. (2012). Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. *Genetics, Selection, Evolution*, 44, 12. https://doi.org/10.1186/1297-9686-44-12

Hughes-Oliver, J. M. (2006). Pooling experiments for blood screening and drug discovery. In A. Dean & S. Lewis (Eds.), *Screening: Methods for experimentation in industry, drug discovery, and genetics* (pp. 48–68). Springer.

Jannink, J. L., Lorenz, A. J., & IWATA, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, 9, 166–177. https://doi.org/10.1093/bfgp/elq001

Lander, E. S., & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185–199. https://doi.org/10.1093/genetics/121.1.185

Li, Y., Levan, O., Kim, J., Zhang, T., Chen, X., & Suo, C. (2019). Extreme sampling design in genetic association mapping of quantitative trait loci using balanced and unbalanced case-control samples. *Scientific Reports*, 9, 15504. https://doi.org/10.1038/s41598-019-51790-w

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

Lovell, J. T., Macqueen, A. H., Mamidi, S., Bonnette, J., Jenkins, J., Napier, J. D., Sreedasyam, A., Healey, A., Session, A., Shu, S., Barry, K., Bonos, S., Boston, L., Daum, C., Deshpande, S., Ewing, A., Grabowski, P. P., Haque, T., Harrison, M., ... Schmutz, J. (2021). Genomic mechanisms of climate adaptation in polypliod bioenergy switchgrass. *Nature*, 590, 438–444. https://doi.org/10.1038/s41586-020-03127-1

Lush, J. L. (1943). *Animal breeding plans* (2nd ed.). Iowa State Collegiate Press, Inc. https://archive.org/details/animalbreedingplan032391mbp/page/n9/mode/2up

McManah, C. S., Tebbs, J. M., & Bilder, C. R. (2012). Two-dimensional informative array testing. *Biometrics*, 68, 793–804. https://doi.org/10.1111/j.1541-0420.2011.01726.x

McManah, C. S., Tebbs, J. M., Hanson, T. E., & Bilder, C. R. (2017). Bayesian regression for group testing data. *Biometrics*, 73, 1443–1452. https://doi.org/10.1111/biom.12704

Meuwissen, T. H. E., Hayes, B. J., & GODDARD, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genet- ics*, 157, 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Moehring, J., Williams, E. R., & Pielpho, H. P. (2014). Efficiency of augmented p-rep designs in multi-environmental trials. *Theoret- ical and Applied Genetics*, 127, 1049–1060. https://doi.org/10.1007/s00122-014-2278-y

Morton, R., & Howarth, J. M. (2005). Estimation of breeding values using selected pedigree records. *Biometrics*, 61, 422–432. https://doi.org/10.1111/j.1541-0420.2005.04035.x

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., & Gerstein, M. (2016). The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biology*, 17, 53. https://doi.org/10.1186/s12359-016-0917-0

Neyhart, J. L., Tiede, T., Lorenz, A. J., & Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3 Genes, Genomes, Genetics*, 7, 1499–1510. https://doi.org/10.1534/g3.117.040550

Okada, M., Lanzatella, C., Saha, M. C., Bouton, J., Wu, R., & Tobias, C. M. (2010). Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. *Genet- ics*, 185, 745–760. https://doi.org/10.1534/genetics.110.113910

Perrin, R. K., Fulginiti, L. E., & Alhassan, M. (2017). Biomass from marginal cropland: Willingness of North Central US farmers to produce switchgrass on their least productive fields. *Biofuels, Bioproducts and Biorefining*, 11, 281–294. https://doi.org/10.1002/bbb.1741

Perrin, R., Vogel, K., Schmer, M., & Mitchell, R. (2008). Farm-scale production cost of switchgrass for biomass. *BioEnergy Research*, 1, 91–97. https://doi.org/10.1007/s12155-008-9005-y

Poudel, H. P., Lee, D., & Casler, M. D. (2020). Selection for winter survivorship in lowland switchgrass. *BioEnergy Research*, 13, 109–119. https://doi.org/10.1007/s12155-020-10091-1

Poudel, H. P., Sanciangco, M. D., Kaeppler, S. M., Buell, C. R., & Casler, M. D. (2019). Genomic prediction for winter survival of lowland switchgrass in the northern USA. *G3 Genes, Genomes, Genetics*, 9, 1921–1931. https://doi.org/10.1534/g3.119.400094

Price, D. L., & Casler, M. D. (2014). Divergent selection for secondary traits in upland tetraploid switchgrass and effects on sward biomass yield. *Bioenergy Research*, 7, 329–337. https://doi.org/10.1007/s12155-013-9374-8

Ramstein, G. P., Evans, J., Kaeppler, S. M., Mitchell, R. B., Vogel, K. P., Buell, C. R., & Casler, M. D. (2016). Accuracy of genomic prediction in switchgrass (*Panicum virgatum* L.) improved by accounting for linkage disequilibrium. *G3 Genes, Genomes, Genetics*, 6, 1049–1062. https://doi.org/10.1534/g3.115.024950

Simeão Resende, R. M., Casler, M. D., & De Resende, M. D. V. (2014). Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree. *Journal of Animal Science*, 92, 4096–4108. https://doi.org/10.2527/jas.2016-0675

Ribaut, J. M., & Ragot, M. (2019). Modernising breeding for orphan crops: Tools, methodologies, and beyond. *Planta*, 250, 971–977. https://doi.org/10.1007/s00425-019-03020-8

Ribaut, J. M., De Vicente, M., & Delannay, X. (2010). Molecular breeding in developing countries: Challenges and perspectives. *Current Opinion in Plant Biology*, 13, 213–218. https://doi.org/10.1016/j.pbi.2009.12.011

Sanchez, D. L., Johnson, N., Mccoy, S. T., Turner, P. A., & Mach, K. J. (2018). Near-term deployment of carbon capture and sequestration from biorefineries in the United States. *Proceedings of the National
Academy of Sciences, 115, 4875–4880. https://doi.org/10.1073/pnas.1719695115

Technow, F., & Gerke, J. (2017). Parent–progeny imputation from pooled samples for cost-efficient genotyping in plant breeding. PLoS ONE, 12, e0190271. https://doi.org/10.1371/journal.pone.0190271

Vogel, K. P., Mitchell, R. B., Casler, M. D., & Sarath, G. (2014). Registration of ‘Liberty’ switchgrass. Journal of Plant Registrations, 8, 242–247. https://doi.org/10.3198/jpr2013.12.0076crc

Werner, C. R., Voss-Fels, K. P., Miller, C. N., Qian, W., Hua, W., Guan, C. Y., Snowdon, R. J., & Qian, L. (2018). Effective genomic selection in a narrow-genepool crop with low-density markers: Asian rapeseed as an example. The Plant Genome, 11, 170084. https://doi.org/10.3835/plantgenome2017.09.0084

Westreich, D. J., Hudgens, M. G., Fiscus, S. A., & Pilcher, C. D. (2008). Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. Journal of Clinical Microbiology, 46, 1785–1792. https://doi.org/10.1128/JCM.00787-07

Williams, E., Piepho, H. P., & Whitaker, D. (2011). Augmented p-rep designs. Biometrical Journal, 53, 19–27. https://doi.org/10.1002/bimj.201000102

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Frontiers in Genetics, 10, 189. https://doi.org/10.3389/fgene.2019.00189

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Tilhou, N. W., & Casler, M. D. (2021). Subsampling and DNA pooling can increase gains through genomic selection in switchgrass. Plant Genome. 2021;14:e20149. https://doi.org/10.1002/tpg2.20149