Genomic Prediction of Biomass Yield in Two Selection Cycles of a Tetraploid Alfalfa Breeding Population

Xuehui Li, Yanling Wei, Ananta Acharya, Julie L. Hansen, Jamie L. Crawford, Donald R. Viands, Réal Michaud, Annie Claessens, and E. Charles Brummer*

Abstract
Alfalfa (Medicago sativa L.) is a widely planted perennial forage legume grown throughout temperate and dry subtropical regions in the world. Long breeding cycles limit genetic improvement of alfalfa, particularly for complex traits such as biomass yield. Genomic selection (GS), based on predicted breeding values obtained using genome-wide molecular markers, could enhance breeding efficiency in terms of gain per unit time and cost. In this study, we genotyped tetraploid alfalfa plants that had previously been evaluated for yield during two cycles of phenotypic selection using genotyping-by-sequencing (GBS). We then developed prediction equations using yield data from three locations. Approximately 10,000 single nucleotide polymorphism (SNP) markers were used for GS modeling. The genomic prediction accuracy of total biomass yield ranged from 0.34 to 0.51 for the Cycle 0 population and from 0.21 to 0.66 for the Cycle 1 population, depending on the location. The GS model developed using Cycle 0 as the training population in predicting total biomass yield in Cycle 1 resulted in accuracies up to 0.40. Both genotype × environment interaction and the number of harvests and years used to generate yield phenotypes had effects on prediction accuracy across generations and locations. Based on our results, the selection efficiency per unit time for GS is higher than phenotypic selection, although accuracies will likely decline across multiple selection cycles. This study provided evidence that GS can accelerate genetic gain in alfalfa for biomass yield.

Alfalfa is one of the most important perennial forage legumes in temperate and subtropical regions in the world. Most alfalfa cultivars are synthetic populations of genetically heterogeneous plants that are heterozygous at many loci throughout the genome. Breeding is conducted as recurrent selection with or without progeny testing. The focus of most breeding programs is on improvements in disease and insect resistance, which have been striking over the past 60 yr, on long-term persistence and on other specific traits such as multifoliolate leaflets and, recently, transgenes such as Roundup Ready. Increased yield is often not explicitly selected and, in fact, yield improvement has stagnated in recent years (Brummer and Casler, 2014). Even with intentional selection for biomass yield, progress is likely to be slow due to long selection cycles required to adequately assess multiyear persistence.

One way to improve breeding selection efficiency is to select molecular markers associated with desirable target trait alleles (Lande and Thompson, 1990). Genomic selection using markers across the genome to predict breeding values could accelerate genetic gain in...
both animals and plants. To conduct GS, phenotypic and genotypic data are collected in a training population so that effects of all markers can be simultaneously estimated to develop a prediction equation (Meuwissen et al., 2001). Based on the GS prediction model, genomic estimated breeding values (GEBVs) are computed on individual plants and serve as the basis for selection in subsequent cycles. Similar to marker-assisted selection (MAS), GS could accelerate breeding by selecting candidate individuals at an early stage or in an off-season nursery or green house. Instead of only focusing on previously identified major quantitative trait loci (QTL) in MAS, small- to medium-effect QTL could also be captured using markers across the entire genome, thereby contributing to prediction accuracy in GS (Meuwissen et al., 2001). Numerous simulation and empirical studies have shown that GS is superior to MAS on prediction accuracy (Bernardo and Yu, 2007; Heffner et al., 2011; Lorenzana and Bernardo, 2009; Zhong et al., 2009). Genomic selection has been reported to result in greater response in actual selection experiments and breeding programs (Asoro et al., 2013; Massman et al., 2013).

Alfalfa and other perennial crops could greatly benefit from GS by reducing the time to complete each breeding cycle (Hayes et al., 2013; Li and Brummer, 2012; Resende et al., 2014). Transcriptome sequencing has identified millions of SNPs in alfalfa (Han et al., 2011; Li et al., 2012; Yang et al., 2011), some of which have been used to develop an Illumina Infinium SNP array containing about 10,000 SNP markers (Li et al., 2014a). In addition, we have identified thousands of SNP using GBS (Elshire et al., 2011) and have genetically mapped some of a tetraploid alfalfa biparental population (Li et al., 2011) and have genetically mapped some of a tetraploid alfalfa biparental population (Li et al., 2014b). These advances in genotyping make exploration of GS in an alfalfa breeding population possible.

For a previous experiment, we conducted two cycles of recurrent selection for high yield in the autotetraploid alfalfa breeding population NY0358. In this experiment, we genotyped the same plants that were evaluated in each of those two selection cycles using GBS and explored genome prediction accuracy of biomass yield across locations and breeding cycles.

**MATERIALS AND METHODS**

**Plant Material**

The tetraploid breeding population (NY0358) used in this experiment was described in Li et al. (2011b). Briefly, NY0358 was a strain cross between three commercial alfalfa cultivars, 5454, Oneida VR (Viands et al., 1990), and AC Viva. The fall dormancy ratings (Teuber et al., 1998) were 4 for 5454 and 3 for Oneida VR and AC Viva. Approximately 100 plants of each cultivar were intercrossed in the greenhouse. Seeds from those crosses were germinated in the greenhouse, and approximately 100 randomly chosen progeny were subsequently intercrossed to form the Cycle 0 base population.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Location</th>
<th>Year</th>
<th>Harvest</th>
<th>Abbreviation</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle 0</td>
<td>Ames, Iowa</td>
<td>2005</td>
<td>1</td>
<td>IA05-1</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>IA05-2</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>IA05-3</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>IA05-4</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2006</td>
<td>1</td>
<td>IA06-1</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>IA06-2</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Ithaca, New York</td>
<td>2005</td>
<td>1</td>
<td>NY05-1</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>NY05-2</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>NY05-3</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2006</td>
<td>1</td>
<td>NY06-1</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>NY06-2</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>NY06-3</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Levis, Quebec</td>
<td>2006</td>
<td>1</td>
<td>QC06-1</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>QC06-2</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Ithaca, New York</td>
<td>2010</td>
<td>1</td>
<td>NY10-1</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>NY10-2</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011</td>
<td>1</td>
<td>NY11-1</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>NY11-2</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>NY11-3</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Levis, Quebec</td>
<td>2010</td>
<td>1</td>
<td>QC10-1</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>QC10-2</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>QC10-3</td>
<td>0.85</td>
</tr>
</tbody>
</table>

**Phenotypic Selection and Data Collection**

Evaluations were conducted using clonal ramets in each of two cycles of selection. Details of the first cycle of selection were given in Li et al. (2011b). Briefly, cuttings of 190 randomly selected genotypes from the Cycle 0 population were made in the greenhouse. The rooted clones were then planted into field experiments at the Agronomy and Agricultural Engineering Research Farm west of Ames, IA, in April 2004; at the Snyder 5 east field adjacent to the Game Farm Road Weather Station in Ithaca, NY, in June 2004; and at the Harlaka experimental site in Levis, QC, in May 2005. At each location, each genotype was clonally replicated nine times in the experiment, with three clones of each genotype representing an experimental unit in each of three replications of a randomized complete block design. Clones were spaced 15 cm apart within plot and plots were spaced 75 cm apart with rows. Rows were 75 cm apart. Nurseries in Iowa and New York were oversown with the low-growing turfgrass creeping red fescue (Festuca rubra L.) to minimize weed encroachment and eliminate the need for cultivation. No yield data were collected in the year of establishment. Yield data were collected in 2005 at Iowa and New York and at all locations in 2006 by hand clipping forage approximately 5 cm above the soil surface. In New York and Iowa, grass was removed before weighing each plot. Between two and four harvests were taken at each location in each year (Table 1). For each harvest in a given year and location, random samples were taken of harvested biomass, weighed wet, dried for 5 d at 60°C
in a forced-air dryer, and then weighed dry. The average dry-matter content of the samples was used to compute dry-matter biomass yield of each plant.

The 20 highest-yielding genotypes, based on the mean biomass yield across the three locations, were selected from Cycle 0 and randomly mated to generate the Cycle 1 population. In spring 2009, 185 Cycle 1 genotypes were clonally propagated and planted at New York and Quebec as described above; Iowa was not included in the second cycle. Biomass yield was measured at New York in 2010 (NY10) and 2011 (NY11) and at Quebec in 2010 (QC10) (Table 1). Two or three harvests were taken each year at each location (Table 1). The dry-matter biomass yield was collected for each harvest as described in Cycle 0 above, with the exception of the first harvest at New York in 2011. For that harvest, grass was mowed in between rows before harvest, but grass within the row was included in the plot weight.

Phenotypic Data Analysis

Yield data were collected for 14 harvests across three locations for Cycle 0 and eight harvests across two locations for Cycle 1. The yield of individual harvests at a location was summed to obtain total biomass yield at each location. Total biomass yield across locations was then analyzed using a generalized linear model as follows:

\[ y_{ijk} = m + l_{ik} + r_{ijk} + g_{ij} + e_{ijk} \]  

[1]

in which \( y_{ijk} \) is the yield for the \( j \)th genotype in the \( i \)th replication of the \( k \)th location, \( m \) is the grand mean, \( l_{ik} \) is the effect of the \( k \)th location, \( r_{ijk} \) is the effect of \( i \)th replication nested in the \( k \)th location, \( g_{ij} \) is the genetic effect of the \( j \)th genotype, \( e_{ijk} \) is the residual. All factors were considered as random, and the ANOVA was performed using PROC GLM (SAS Institute, 2010). Significant genotype \( \times \) location interaction was found in both cycles (data not shown), and therefore, we subsequently analyzed individual locations separately.

For each individual harvest in a given environment, best linear unbiased predictors (BLUPs) were estimated using a random effects model:

\[ y_{j} = m + r_{j} + g_{j} + e_{j} \]  

[2]

in which \( y_{j} \) is the yield value for the \( j \)th genotype in the \( i \)th replication, \( m \) is the grand mean yield of the specific harvest in the specific environment being evaluated, \( r_{j} \) is the effect of the \( i \)th replication, \( g_{j} \) is the genetic effect of the \( j \)th genotype, and \( e_{j} \) is the residual. The grand mean was the only fixed effect and others were considered as random effects. The BLUPs of total biomass yield at each location were estimated with the same mixed model (Eq. [2]), substituting total biomass yield for yield of a specific harvest. The estimated BLUPs, denoted as \( y_{j} \), were considered as the observed phenotypic values for GS prediction model evaluation.

To estimate the BLUP of total biomass yield across locations, the estimated BLUPs at each location were fit to a mixed model as follows:

\[ y_{jk} = m + l_{k} + g_{j} + e_{jk} \]  

[3]

in which \( y_{jk} \) is the estimated BLUPs for the \( j \)th genotype in the \( k \)th location, \( m \) is the grand mean, \( l_{k} \) is the effect of the \( k \)th location, \( g_{j} \) is the genetic effect of the \( j \)th genotype. The grand mean and location were considered as fixed effects and the genotypes were treated as random effects.

In all cases, the BLUPs estimation was performed using the R package lme4 (R Development Core Team, 2012; Bates et al, 2011). The estimated BLUPs were considered as the observed phenotypic values of total biomass yield for further GS prediction model evaluation.

The genotype (\( \sigma_{g}^{2} \)) and residual (\( \sigma_{e}^{2} \)) variance components were estimated using the mixed model (Eq. [2]) and used to compute broad-sense heritability as follows:

\[ H^{2} = \frac{\sigma_{g}^{2}}{\sigma_{e}^{2} + \sigma_{g}^{2}} \]  

[4]

DNA Isolation, Genotyping-by-Sequencing

Library Construction, and Sequencing

DNA was isolated from Cycle 0 plants using the traditional CTAB method (Doyle and Doyle, 1990) in a 96-well format; for Cycle 1, DNA was isolated with the Wizard Genomic DNA Purification Kit (Promega, A1125) per the manufacturer’s instructions. DNA was quantified with a Quant-iT PicoGreen dsDNA assay kit (Life Technologies, P7589). One library was constructed each for Cycle 0 at 190-plex and Cycle 1 at 185-plex, using the protocol of Elshire et al. (2011) with minor modifications. Briefly, 100 ng of each DNA was digested with ApeKI (NEB, R0643L) and then ligated to a unique barcoded adaptor and a common adaptor. An equal volume of the ligated product from each genotype was pooled and the mixture was purified using the QIAquick PCR purification kit (QIAGEN, 28104) before PCR amplification. In the PCR, 50 ng of template DNA was mixed with NEB 2X Taq Master Mix and two primers (5 nmoles each) in a 50 μL total volume and amplified on a thermocycler for 18 cycles with 10 s of denaturation at 98°C, followed by 30 s of annealing at 65°C and, finally, 30 s extension at 72°C. Each library was sequenced in two lanes on Illumina HiSeq 2000 at the Genomic Sequencing and Analysis Facility at the University of Texas at Austin, Texas, USA. All sequences were submitted to the National Center for Biotechnology Information (NCBI) Short Read Archive (study #SRX685967).
Genotype Calling
The UNEAK pipeline (Lu et al., 2013) was used for SNP discovery and genotype calling. Briefly, the raw reads (100 bp, single-end read) obtained from the sequencer were first quality filtered and demultiplexed. All reads that began with one of the expected barcodes immediately followed by the expected cut site remnant (CAGC or CTGC for ApeKI) were trimmed to 64 bp (including the cut site remnant but removing the barcode). Reads with identical sequences were grouped into one tag. The tags with 10 or more reads across all individuals were retained for pairwise alignment. Pairwise alignment was performed to find pairs of tags that differed by only 1 bp. For each SNP marker, the read distribution of the paired tags in each individual was used for SNP genotype calling as described by Li et al. (2014b). Briefly, for a given SNP (e.g., A/T), if only a single allele was observed for a given individual, then a minimum of 11 reads was required to call a homozygote (e.g., AAAA). If fewer than 11 reads were present, we assigned a missing genotype to avoid misclassifying a triplex heterozygote (AAAT) as homozygous. The probability of miscalling a triplex heterozygote as a homozygote (AAAA) is <0.05 if 11 or more reads are present and allele amplification is unbiased. When both alleles were observed in a given individual, we required a minimum of two reads per allele and a minimum minor allele frequency for that locus in a given individual greater than 0.10 to call a heterozygote; otherwise, a missing genotype call (NA) was assigned. Requiring two reads of the minor allele limits the likelihood that an allele resulted from a sequencing error. However, if a large number of sequencing reads are available for a given locus, multiple sequencing errors might be likely. Therefore, we included the minor allele frequency limit to avoid calling homozygotes as heterozygotes that were obtained solely due to sequence errors. Reliably discriminating among the three heterozygote genotypes in an autotetraploid would require a read depth of at least 60 (Uitdewilligen et al., 2013). Because only a small percentage of our GBS SNP markers met this criterion, we did not attempt to distinguish among heterozygote genotypes. The Basic Local Alignment Search Tool (BLAST) was used to query the consensus sequence of each tag pair containing a SNP against the *Medicago truncatula* reference genome Version 4.1. The best hit with a cutoff of E-value <1 × 10^{-5} was selected as the physical location of the GBS SNP markers.

Population Structure and Linkage Disequilibrium Estimation
For each breeding cycle, the GBS SNP markers with fewer than 20% missing values were used for model-based clustering using the R package HDclassif (Berge et al., 2012). The best model, which had the optimal number of clusters, was established according to the Bayesian information criterion (BIC).

For each cycle, the linkage disequilibrium (LD) between pairs of GBS SNP markers was estimated as $r^2$ using the software program TASSEL (Bradbury et al., 2007) with heterozygous calls treated as missing values. We estimated LD in four datasets representing GBS SNP markers with fewer than 30, 50, 70, or 80% missing genotypic values and, in all cases, with a minor allele frequency higher than 5%. The functional relationship between LD and physical distances on the *M. truncatula* genome was evaluated by fitting a nonlinear model. The expected $r^2$ under drift-recombination equilibrium was $E(r^2) = 1/(1 + C)$ and $C = 4ad$, where $d$ is the physical distance in bp and $a$ is a regression coefficient estimated from our data (Sved 1971). With a low level of mutation and finite sample size $n$, the expectation becomes the following (Hill and Weir 1988):

$$E(r^2) = \left[\frac{10+C}{2+G(1+G)}\right]^2 + \left[\frac{3+C}{2+G(1+G)}\right]^2 \left[\frac{1}{n} + \frac{1}{2+G(1+G)}\right]$$  \[5\]

Genomic Selection Prediction Model Development and Validation
The general GS prediction model used was the following:

$$y = m + Zu + e$$  \[6\]

in which $y$ is the observed phenotypic value (estimated BLUP from Eq. [1]), $m$ is the intercept, $Z$ is the marker matrix, and $u$ is a vector of estimated marker effects. Ridge regression method was used to estimate effects of markers using the R package rrBLUP (Endelman, 2011). The GBS SNP marker data sets with varying levels of missing values were used to model each trait at each environment. Before fitting the models, the missing values were imputed using the random forest method with the R package missForest (Stekhoven and Buhlmann, 2012).

We evaluated genomic prediction accuracy for single-harvest biomass yield in each location and total biomass yield at each location and across locations. The genomic prediction model was validated by cross validation, in which 90% of individuals were randomly selected as a training population to estimate marker effects and the remaining 10% of individuals were used to validate the genomic prediction accuracy. The genomic prediction accuracy was estimated as the Pearson correlation ($r$) between estimated GEBVs and estimated BLUPs of phenotypic values. Random sampling training and validation sets were repeated 3000 times and the mean of correlations was defined as the genomic prediction accuracy.

For total biomass yield, we evaluated the accuracy of GS models developed using one location as the training population to predict biomass yield in other locations. We also evaluated the accuracy of GS models developed using Cycle 0 as the training population to predict biomass yield in Cycle 1. For this analysis, all individuals from Cycle 0 were used to develop a GS prediction model and all individuals from Cycle 1 were used for validation based on the SNP markers in common with the Cycle 0 training set.
RESULTS

Phenotype Data
The broad-sense heritability of biomass yield ranged from 0.70 to 0.93 across harvests for Cycle 0 and from 0.61 to 0.88 across harvests for Cycle 1 (Table 1). These results indicated that large genetic variation for biomass yield existed in the population at all harvests and in all environments in both Cycle 0 and Cycle 1.

Genotyping-by-Sequencing Genotype Data
After quality filtering and processing, we identified a total of 448,080,472 sequencing reads for Cycle 0 and 369,843,942 for Cycle 1. The read number per genotype was 2,358,318 on average for Cycle 0, ranging from 430,090 to 10,390,898. One genotype in Cycle 1 had only 2776 reads and was dropped from further analysis. The average read number per genotype in Cycle 1 was 2,010,006, ranging from 1,213,037 to 6,354,979.

The sequence reads from both breeding cycles were analyzed together using the UNEAK pipeline. In total, 177,205 GBS SNP markers were identified in this breeding population and 73,256 of them (41.3%) aligned to the eight chromosomes of \( M. \) truncatula pseudomolecules v4.1. After read-depth filtering, we genotyped 18,525 SNPs in Cycle 0 and 14,913 in Cycle 1 that had up to 90% missing values (Supplemental Table S1, S2). We aligned 76% of Cycle 0 SNPs and 77% of Cycle 1 SNPs to the eight chromosomes of the \( M. \) truncatula reference genome (Supplemental Table S1, S2). Within these SNP datasets, 7075 SNPs in Cycle 0 and 6241 SNPs in Cycle 1 had fewer than 50% missing values and ~85% aligned to the \( M. \) truncatula pseudomolecules (Supplemental Table S1, S2). For the markers having fewer than 10% missing values, about 90% aligned to the \( M. \) truncatula pseudomolecules (Supplemental Table S1, S2). In all cases, we imputed missing data using random forest imputation before model development.

Population Structure and Linkage Disequilibrium
For each breeding cycle, we used model-based clustering analysis of GBS SNP markers with fewer than 20% missing values to assess population substructure. Based on the BIC, the best model for either Cycle 0 or Cycle 1 was a single cluster (data not shown), suggesting no subpopulation structure.

We measured LD as \( r^2 \) between a pair of markers based on marker data sets having missing values for fewer than 30, 50, 70, or 80% of the genotypes. The estimated LD decreased as the percentage of missing values increased to 70% in both cycles (Fig. 1). Using the dataset with SNP markers having fewer than 70% missing values, LD in Cycle 0 decayed to 0.42 at 200 Kbp between markers, but in Cycle 1, LD was more extensive, with \( r^2 = 0.74 \) at 200 Kbp (Fig. 1).
Genomic Prediction Accuracy

Marker number is one of major factors affecting genomic prediction accuracy (Heffner et al., 2011; Heslot et al., 2013; Poland et al., 2012; Zhong et al., 2009). For single-harvest biomass yield at each location and total biomass yield at each location and across locations, we developed and validated GS models using a series of GBS SNP marker datasets that had a threshold for missing values that ranged from 10 to 90% of the genotypes in the population. Marker sets that accepted higher amounts of missing data had more markers. We found that the prediction accuracies increased as missing values increased to between 60 and 70% in Cycle 0 (Supplemental Table S3) and to 80% in Cycle 1 (Supplemental Table S4). The harvest NY11-1 in Cycle 1 had consistently low accuracies and was unusual compared with all other harvests (Supplemental Table S4). The broad-sense heritability of this particular harvest was much lower than other harvests, implying poor quality phenotypic data (Table 1). These results likely resulted because a flail-type harvester was used for this harvest and the oversown companion grass was included in the plot weight. Based on these results, we used SNP datasets represented by up to 65% missing data for Cycle 0 and up to 80% missing data for Cycle 1 to develop and validate GS models.

Validation within Cycle 0

Using the dataset of 9906 markers that each had fewer than 65% missing genotypic values, the prediction accuracy among the 14 harvests in Cycle 0 was 0.26 to 0.51 (Supplemental Table S3). Accuracies from one harvest from Iowa (IA05-1) and one harvest from New York (NY06-3) were lower than 0.30 (Supplemental Table S3).

In Cycle 0, six harvests in total from 2 yr were taken at Iowa and New York and two harvests from 1 yr at Quebec. We further evaluated genomic prediction accuracy of total biomass yield within a location and between locations. The prediction accuracy of total biomass yield was 0.43 at Iowa, 0.49 at New York, 0.44 at Quebec, and 0.51 across the three locations (Table 2). The prediction accuracies validated between locations, especially between Iowa and Quebec, were much lower than prediction accuracies validated within location (Table 2). Based on New York data only, 16 of the top 20 highest yielding genotypes would have also been selected using GEBVs (data not shown). Across all locations, 11 of 20 individuals selected for maximum yield based on phenotypic data were also among the highest 20 individuals based on GEBVs (data not shown).

Validation within Cycle 1

In Cycle 1, an accuracy of 0.40 to 0.67 was observed for the seven individual harvests (excluding NY11-1) in New York and Quebec using the 11,282 markers with fewer than 80% missing values (Supplemental Table S4). The prediction accuracy of total biomass yield was 0.45 at New York based on all five harvests and 0.51 based on the four harvests excluding NY11-1 (Table 3). The prediction accuracy of total biomass yield was 0.66 for Quebec and 0.59 across the two locations (Table 3). The prediction accuracies validated between the two locations were two to three times lower than the prediction accuracies validated within location (Table 3).

Prediction across Generations

We further evaluated prediction accuracy of total biomass yield across generations. We identified 9077 of the 9906 markers that were part of the Cycle 0 data set that were also part of the Cycle 1 dataset. We developed a genomic prediction model considering Cycle 0 as the training population to then predict biomass yield in Cycle 1. The GS model developed from New York Cycle 0 showed an accuracy of 0.40 when applied to data from New York Cycle 1, the best of all models, and 0.39 when applied to Quebec Cycle 1 (Table 4). Models developed from Quebec Cycle 0 had low prediction accuracy for Cycle 1 at either location. Iowa Cycle 0 models also had low prediction for Cycle 1 from New York or Quebec (Table 4). Models developed from total yield across the three locations at Cycle 0 showed higher prediction accuracies than Iowa and Quebec Cycle 0 models, but lower prediction accuracies relative to the New York Cycle 0 model (Table 4). Removing Iowa and analyzing New York and Quebec together resulted in an accuracy of 0.38 for predicting mean yield across New York and Quebec in the Cycle 1 population.

### Table 2. The genomic prediction accuracy of total biomass yield in Cycle 0 validated within and across locations.

<table>
<thead>
<tr>
<th>Estimation</th>
<th>IA</th>
<th>NY</th>
<th>QC</th>
<th>IA-NY-QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>0.43</td>
<td>0.43</td>
<td>0.34</td>
<td>0.47</td>
</tr>
<tr>
<td>NY</td>
<td>0.37</td>
<td>0.49</td>
<td>0.39</td>
<td>0.48</td>
</tr>
<tr>
<td>QC</td>
<td>0.34</td>
<td>0.46</td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td>IA-NY-QC</td>
<td>0.40</td>
<td>0.49</td>
<td>0.41</td>
<td>0.51</td>
</tr>
</tbody>
</table>

*1 IA, Ames, IA; NY, Ithaca, NY; QC, Levis, QC; IA-NY-QC is the total biomass yield across the three locations of IA, NY, and QC.

### Table 3. The genomic prediction accuracy of total biomass yield in Cycle 1 validated within and across locations.

<table>
<thead>
<tr>
<th>Estimation</th>
<th>NY-A</th>
<th>NY-B</th>
<th>QC</th>
<th>NY-QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NY-A</td>
<td>0.45</td>
<td>0.49</td>
<td>0.22</td>
<td>0.29</td>
</tr>
<tr>
<td>NY-B</td>
<td>0.47</td>
<td>0.51</td>
<td>0.21</td>
<td>0.28</td>
</tr>
<tr>
<td>QC</td>
<td>0.22</td>
<td>0.21</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>NY-QC</td>
<td>0.29</td>
<td>0.29</td>
<td>0.61</td>
<td>0.59</td>
</tr>
</tbody>
</table>

*1 NY, Ithaca, NY; QC, Levis, QC; NY-A is the total biomass yield from all five harvests at NY; NY-B is the total biomass yield from the four harvests by excluding the harvest NY11-1; NY-QC is the total biomass yield across the two locations of NY and QC.*
The number of markers needed to perform GWAS or GS in this population can be inferred from LD estimates. Linkage disequilibrium quickly decayed within a few Kbp in a wild diploid alfalfa collection based on SSR markers (Sakiroglu et al., 2012). Using the Illumina Infinium alfalfa SNP array, we evaluated genome-wide LD with a large number of markers and found that LD quickly decayed to 26 Kbp at $r^2 = 0.2$ over all germplasms tested but varied by population (Li et al., 2014a). In this study, we estimated LD of the two breeding cycles using GBS SNP markers and found that the estimated LD decreased as the percentage of missing values increased to 70%. Removing markers with more missing values results in fewer markers tested, which could bias the LD estimation. In contrast, including the markers with more missing values results in a small sample size for some markers, which could cause an overestimation of LD (Yan et al., 2009). Using markers with fewer than 70% missing values, we found that the LD is about $r^2 = 0.42$ in Cycle 0 and $r^2 = 0.74$ in Cycle 1 at a distance of 200 Kbp.

One cycle of selection likely caused the increased LD in Cycle 1. Based on the estimated LD, about 42% of the genetic variation in Cycle 0 could be captured by using 5000 markers assuming an alfalfa genome size of 1000 Mbp and equally distributed markers throughout the genome. However, our LD estimates were made based on the physical distances between markers in the M. truncatula reference genome. Although numerous studies have shown a high level of synteny between M. truncatula and alfalfa (Choi et al., 2004; Li et al., 2011a), the estimated LD could be biased because we do not have a full genome sequence of alfalfa for comparison. Nevertheless, while we did not completely cover the genome with this experiment, we were able to make considerable improvement relative to previous mapping experiments.

### Table 4. Genomic prediction accuracy of total biomass yield across generations.

<table>
<thead>
<tr>
<th>Estimation in Cycle 0</th>
<th>Validation in Cycle 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NY-A</td>
</tr>
<tr>
<td>IA</td>
<td>0.26</td>
</tr>
<tr>
<td>NY</td>
<td>0.40</td>
</tr>
<tr>
<td>QC</td>
<td>0.26</td>
</tr>
<tr>
<td>NY-QC‡</td>
<td>0.36</td>
</tr>
<tr>
<td>IA-NY-QC</td>
<td>0.38</td>
</tr>
</tbody>
</table>

‡ IA, Ames, IA; NY-QC is the total biomass yield across the two locations of NY and QC; IA-NY-QC is the total biomass yield across the three locations of IA, NY, and QC in Cycle 0.

**DISCUSSION**

**Genotyping-by-Sequencing**

With GBS, we were able to identify over 175,000 SNP in a tetraploid alfalfa breeding program. However, only about 10% of these SNP could be genotyped due to an insufficient read depth in enough individuals in the population for the loci to be useful in mapping. The main issue we face in alfalfa (as in other allogamous polyploids) is heterozygosity; to effectively call a genotype as heterozygous or homozygous, multiple reads are required, and consequently, only 18,525 GBS SNP markers were genotyped in our population. These markers were distributed throughout the M. truncatula reference genome. About 90% of loci for which genotype calls could be made in over 90% of individuals aligned to the M. truncatula reference genome, but only 41.3% of all markers could be aligned. Genomic regions with highly conserved sequences between M. truncatula and M. sativa—such as exons—would also be more conserved among diverse alfalfa haplotypes and therefore, possibly more commonly sequenced.

Large numbers of genome-wide markers are needed to apply genomic selection to alfalfa or to conduct various other mapping and trait dissection experiments. Previously, we developed an Illumina Infinium array with ~10,000 features (Li et al., 2014a). Arrays have particular advantages, including very little missing data, known loci as markers (selected for specific reasons), and facile comparison among experiments. However, they suffer ascertainment bias, and the number of markers is fixed, so that experiments needing fewer markers or those requiring many more cannot be accommodated, and the cost is quite high on an experiment basis. In contrast, GBS-based SNP markers are typically cheaper, and assuming a relatively large amount of missing data can be handled and accepted, they represent an alternative high-throughput genotyping platform. Compared to the SNP array, GBS has no ascertainment bias and is tunable to generate more or fewer markers depending on the purpose of the experiment. Using GBS, we have generated about 9000 SNP markers polymorphic for a tetraploid alfalfa biparental F₁ mapping population and mapped over 3500 SNP markers on 32 linkage groups for each of the two parents, four linkage groups per chromosome representing the four haplotypes (Li et al., 2014b). For our purposes here, GBS appears to be the only way to get sufficient markers within budget constraints to implement a GS program in alfalfa at the current time.

**Population Structure and Linkage Disequilibrium**

Population structure affects GS accuracy and needs to be integrated into a GS model if present (Lorenz et al., 2011; Hayes et al., 2009; Toosi et al., 2010). Based on 71 SSR markers, we found no subpopulation structure in Cycle 0 of this breeding population (Li et al., 2011b). Using the GBS SNP markers, we conducted model-based analysis and found no subpopulations in either breeding cycle. These results indicated that the two generations of random mating used to develop Cycle 0 of this breeding population successfully intermixed the parental genomes.

With GBS, we were able to identify over 175,000 SNP in a tetraploid alfalfa breeding program. However, only about 10% of these SNP could be genotyped due to an insufficient read depth in enough individuals in the population for the loci to be useful in mapping. The main issue we face in alfalfa (as in other allogamous polyploids) is heterozygosity; to effectively call a genotype as heterozygous or homozygous, multiple reads are required, and consequently, only 18,525 GBS SNP markers were genotyped in our population. These markers were distributed throughout the M. truncatula reference genome. About 90% of loci for which genotype calls could be made in over 90% of individuals aligned to the M. truncatula reference genome, but only 41.3% of all markers could be aligned. Genomic regions with highly conserved sequences between M. truncatula and M. sativa—such as exons—would also be more conserved among diverse alfalfa haplotypes and therefore, possibly more commonly sequenced.

Large numbers of genome-wide markers are needed to apply genomic selection to alfalfa or to conduct various other mapping and trait dissection experiments. Previously, we developed an Illumina Infinium array with ~10,000 features (Li et al., 2014a). Arrays have particular advantages, including very little missing data, known loci as markers (selected for specific reasons), and facile comparison among experiments. However, they suffer ascertainment bias, and the number of markers is fixed, so that experiments needing fewer markers or those requiring many more cannot be accommodated, and the cost is quite high on an experiment basis. In contrast, GBS-based SNP markers are typically cheaper, and assuming a relatively large amount of missing data can be handled and accepted, they represent an alternative high-throughput genotyping platform. Compared to the SNP array, GBS has no ascertainment bias and is tunable to generate more or fewer markers depending on the purpose of the experiment. Using GBS, we have generated about 9000 SNP markers polymorphic for a tetraploid alfalfa biparental F₁ mapping population and mapped over 3500 SNP markers on 32 linkage groups for each of the two parents, four linkage groups per chromosome representing the four haplotypes (Li et al., 2014b). For our purposes here, GBS appears to be the only way to get sufficient markers within budget constraints to implement a GS program in alfalfa at the current time.

**Population Structure and Linkage Disequilibrium**

Population structure affects GS accuracy and needs to be integrated into a GS model if present (Lorenz et al., 2011; Hayes et al., 2009; Toosi et al., 2010). Based on 71 SSR markers, we found no subpopulation structure in Cycle 0 of this breeding population (Li et al., 2011b). Using the GBS SNP markers, we conducted model-based analysis and found no subpopulations in either breeding cycle. These results indicated that the two generations of random mating used to develop Cycle 0 of this breeding population successfully intermixed the parental genomes.
Genomic Prediction Accuracy and Implication in Alfalfa Breeding

We imputed missing GBS SNP genotypes using the random forest method. Imputation errors are expected to increase as the amount of missing marker data increases (Poland et al., 2012). However, because the number of markers increases in the dataset as the stringency for missing data is relaxed, the imputation errors may be outweighed by improved models with more markers. In this study, we found that the average GS accuracies increased as marker number increased until the allowable amount of missing data per marker exceeded 70 to 80% in spite of the potential higher imputation error caused by more missing values. Our results are congruent with an experiment in wheat (Triticum aestivum L.) where 35,000 GBS SNP markers with up to 80% missing data points gave higher prediction accuracy than 2000 diversity array technology (DArT) markers with 2% missing values (Poland et al., 2012). The better genomic prediction accuracy observed using GBS was mainly due to the large increase in marker numbers and not to ascertainment bias from the DArT markers (Heslot et al., 2013). Thus, more markers, even with increased amounts of missing values, will generally result in higher prediction accuracy until the missing data percentage per marker becomes very high (>70–80%).

High genomic prediction accuracies were observed in numerous empirical studies from a diverse set of plant species such as 0.39 for leaf length and 0.48 for leaf width in maize (Zea mays L.) (Guo et al., 2013), 0.72 for Fusarium head blight resistance in barley (Hordeum vulgare L.) (Lorenz et al., 2012), 0.70 to 0.90 for various fruit quality traits in apple (Malus × domestica Borkh.) (Kumar et al., 2012), and 0.63 to 0.74 for height in loblolly pine (Pinus taeda L.) (Resende et al., 2012). As a perennial forage crop, alfalfa generally survives in the field at least 3 to 4 yr and, depending on location, can be harvested from two to 10 or more times each year. If genomic prediction accuracies of total biomass yield can be identified that are reasonably robust, then the gain from GS could be significant relative to phenotypic selection that occurs after 3 to 4 yr. In this experiment, we observed prediction accuracies of 0.43 to 0.66 for total biomass yield when validated within a location, providing evidence that genomic prediction accuracies in an autotetraploid species and in a synthetic alfalfa breeding population could conceivably be high enough to make GS worthwhile for complex traits like biomass yield. We also found that one harvest (NY11-1) caused lower prediction accuracy when it was included in total biomass yield for that location, reinforcing the point that high quality phenotypic data from as many harvests as possible is essential for alfalfa GS prediction models. Lower prediction accuracies between locations than the accuracies validated within location suggest presence of genotype × location interaction for biomass yield, as found in the ANOVA, particularly between Iowa and the two northeastern North American locations.

A scheme of GS in alfalfa could be developed in which one generation is used as a training population to develop a model that is then applied to selection in later generations (Li and Brummer, 2012). By making use of phenotypic and genotypic data from the two breeding cycles, we estimated the GS prediction accuracy across generations to be 0.40 for populations grown in New York. Even more interesting was the 0.38 prediction accuracy of Cycle 1 New York–Quebec yield based on the New York–Quebec model from Cycle 0. The lower GS accuracy across generations compared with validating within a generation is expected and can be explained by less relatedness of individuals across generations than within generations and by genotype × environment interaction (Ly et al., 2013). In this experiment, the Cycle 1 genotypes are the products of phenotypic selection; we did not conduct GS to generate the Cycle 1 population. The prediction accuracy of the model developed in Cycle 0 could have been different if the Cycle 1 population had been developed based on GS selection.

For the GS model developed at Iowa in Cycle 0, the accuracy in predicting biomass yield in Cycle 1 was low for New York and very poor for Quebec. This implied that there was genotype × environment interaction on biomass yield between the Iowa and Quebec locations. However, for the GS model developed at Quebec in Cycle 0, the accuracy in predicting biomass yield in Cycle 1 was very low even at the same location. Only two harvests were taken from the first production year at Quebec in Cycle 0. No data were collected in subsequent years because many plants were killed by severe winter. This could result in a GS model that did not capture most biomass yield genes across ages or growth seasons and therefore had the poor prediction. Taken together, GS models for alfalfa yield prediction need to be developed using multiple harvests from multiple years and likely need to be applied to a certain target breeding region to avoid bias from genotype × environment interaction. Nevertheless, the model developed across all locations from Cycle 0 to predict Cycle 1 still had an accuracy of 0.33, which would still be useful in a breeding program and which shows that the presence of genotype × environment does not preclude across-location prediction, at least in this population.

Compared to phenotypic selection, the higher selection efficiency from GS is mainly derived from the cumulative response of more selection cycles per unit time (Jannink, 2010). In alfalfa, recurrent selection is the common breeding method, and it generally takes 2 to 5 yr to conduct one cycle of selection. It could take only half a year to conduct one cycle of GS. Given the prediction accuracy of 0.40 and assuming a reduction of 75% per breeding cycle, the selection efficiency per unit time of GS is estimated to be about 60% higher than phenotypic selection. The prediction accuracy was estimated as the correlation between GEBVs and observed phenotypes here. A higher correlation between GEBVs and true breeding values could be expected, particularly if the observed phenotypic variation has a significant
component of dominance genetic variance. Our reasonably good accuracies between generations suggest that, in this population, yield is under substantial additive genetic control. The prediction accuracy observed here with no more than 200 individuals of training population can be higher given a larger training population in alfalfa breeding industry. In addition, an improved statistical model that could integrate allele dosage information might result in a better prediction for polyploid alfalfa.

Alfalfa biomass yield has had a low genetic gain per year and has been stagnant since the 1980s (Brummer 1999; Lamb et al., 2006, Brummer and Casler, 2014). In addition to long breeding cycles, less intense selection is another reason for the slower genetic gain. For alfalfa and other perennial forage crops, breeding populations of small size have been commonly evaluated for selection in each cycle of phenotypic selection. More intense selection in a larger selection population could be performed in GS, which could result in higher selection efficiency than phenotypic selection (Heffner et al., 2009).

CONCLUSIONS

Previously, we used GBS to construct a high-density linkage map for a tetraploid alfalfa \( F_2 \) mapping population. Here we showed that GBS could be used to generate a large number of genome-wide markers for GS studies in a tetraploid alfalfa breeding population. The high genomic prediction accuracy of total biomass yield was observed even across generations in this study, and many ways could be implemented to potentially enhance the prediction accuracy in the future studies or breeding applications. This study demonstrates that GS could accelerate improvement of alfalfa biomass yield and other complex traits by shortening breeding cycles and increasing selection intensity.

Acknowledgments

The population NY0358 was initially developed to evaluate the potential value of clonal selection to improve biomass yield as part of the NE1010 Multistate Cooperative Research Project. This experiment was funded in part by the NE1010 Multistate Cooperative Research Project (DRV and ECB) and by a USDA–DOE Plant Feedstock Genomics for Bioenergy grant to ECB, award #2009-65504-05809. We thank Mark Smith for technical assistance in Iowa.

References


Li, X., Y. Wei, A. Acharya, Q. Jiang, J. Kang, and E.C. Brummer. 2014b. A high-density genetic linkage map of autotetraploid alfalfa (Medicago sativa L.) developed using Genotyping-by-Sequencing is highly syntenous with the M. truncatula genome. G3 (Bethesda) 4:1971–1979. doi:10.1534/g3.114.012245


