Association Mapping of Biomass Yield and Stem Composition in a Tetraploid Alfalfa Breeding Population

Xuehui Li¹, Yanling Wei¹, Kenneth J. Moore, Réal Michaud, Donald R. Viands, Julie L. Hansen, Ananta Acharya, and E. Charles Brummer*

Abstract
Alfalfa (Medicago sativa L.), an important forage crop that is also a potential biofuel crop, has advantages of high yield, high lignocellulose concentration in stems, and has low input costs. In this study, we investigated population structure and linkage disequilibrium (LD) patterns in a tetraploid alfalfa breeding population using genome-wide simple sequence repeat (SSR) markers and identified markers related to yield and cell wall composition by association mapping. No obvious population structure was found in our alfalfa breeding population, which could be due to the relatively narrow genetic base of the founders and/or due to two generations of random mating. We found significant LD (p < 0.001) between 61.5% of SSR marker pairs separated by less than 1 Mbp. The observed large extent of LD could be explained by the effect of bottlenecking and selection or the high mutation rates of SSR markers. Total marker heterozygosity was positively related to biomass yield in each of five environments, but no relationship was noted for stem composition traits. Of a total of 312 nonrare (frequency > 10%) alleles across the 71 SSR markers, 15 showed strong association (p < 0.005) with yield in at least one of five environments, and most of the 15 alleles were identified in multiple environments. Only one allele showed strong association with acid detergent fiber (ADF) and one allele with acid detergent lignin (ADL). Alleles associated with traits could be directly applied in a breeding program using marker-assisted selection. However, based on our estimated LD level, we would need about 1000 markers to explore the whole alfalfa genome for association between markers and traits.

Alfalfa (Medicago sativa L.) is one of the most important forage crops in the world due to its high yield and high nutritive value. In 2008, alfalfa was the third widely planted crop after maize (Zea mays L.) and soybean [Glycine max (L.) Merr.] in the United States, with about 8.5 million ha for hay production (USDA; available at http://www.nass.usda.gov/Statistics_by_Subject/index.php?sector=CROPS [verified 6 Jan. 2011]), with dry hay production of about 63.7 million t and a total value of US$10.8 billion. Alfalfa is a potential biofuel crop, with stems used as energy feedstock and leaves as protein meal for livestock feed (Jung et al., 1994b; Samac et al., 2006). The advantages of alfalfa used as a biofuel crop compared to other crops include its perennial nature, which decreases soil erosion, its nitrogen fixation ability, which reduces input cost, its high yield, and its high lignocellulose concentration in stems (Dien et al., 2006).

Yield is critical for all uses of alfalfa. Unfortunately, improvement in yield of alfalfa has stagnated in recent years (Lamb et al., 2006; Riday and Brummer, 2002). Alfalfa yield has increased about 0.5% per year since 1919, much lower than maize (5.3% per year) (USDA; available at http://www.nass.usda.gov/Statistics_by_
In this experiment, our objective was to test the hypothesis that LD in a tetraploid alfalfa breeding population would extend sufficient distances to enable marker–trait associations to be detected with relatively few markers. Given sufficient LD, we hypothesized we could detect QTL for biomass yield and cell wall composition as measured using neutral detergent fiber (NDF), acid detergent fiber (ADF), and acid detergent lignin (ADL).

**Materials and Methods**

**Mapping Population**

The tetraploid population we used in this study was formed by a strain cross between three commercial cultivars, 5454 developed by Pioneer Hybrid, Inc. (Johnston, IA), in the midwestern United States, Oneida VR developed by Cornell University (Ithaca, NY) (Viands et al., 1990), and AC Viva developed by Ag and Agri-Food Canada, Ste. Foy, QC. The cultivars had fall dormancy ratings (Teuber et al., 1998) of 4 for 5454 and 3 for Oneida VR and AC Viva. The pedigrees of the cultivars are unknown, but they could be assumed to be from the same general semidormant germplasm group. The initial strain cross included about 100 individuals from each cultivar. The resulting population was then randomly mated in the greenhouse for two generations using approximately 100 individuals in each generation, after which 190 individuals were randomly selected, genotyped, and clonally propagated for field evaluation.

**Phenotypic Evaluation**

**Experimental Design**

Field experiments were planted 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 Lévis, QC, in May 2005. The experimental design at each location was a randomized complete block design with three replications, with each plot in each replication containing three clones of each genotype. Clones were spaced 15 cm apart within plots; plots were 75 cm apart within and between rows.

**Yield**

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 harvesting and weighing the fresh mass from the entire plot. Three harvests were taken each year at New York in 2005 (NY05) and 2006 (NY06), four harvests at Iowa in 2005 (IA05), and two harvests in 2006 at Iowa (IA06) and Québec (Que06). For each harvest in a given year and location, samples were randomly taken during the harvest period, weighted 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 plot. The yearly yield was estimated as the sum of all harvests at each
location in a given year. We considered each year by location combination as a separate environment.

**Cell Wall Composition**

Samples from the first (June) harvest in 2006 at Iowa were used for analyses of stem cell wall composition. We sampled a single time point from one location for cell wall composition analysis because genotype × environment and genotype × harvest interactions have been rarely observed for alfalfa composition traits (Sheaffer et al., 1998, 2000). After whole plant samples were dried at 60°C for 4 d, stems were separated from leaves and then ground to a fine powder. Near-infrared reflectance spectroscopy (NIRS) (Windham et al., 1989) was used to analyze composition. Duplicate reflectance (log 1/R) measurements between 400 and 2500 nm at 2-nm intervals were taken using a scanning monochromator (Model 6500; NIRSystems, Silverspring, MD). A subset of 50 samples was selected for calibration on the basis of spectral characteristics. An ANKOM200 Fiber Analyzer (ANKOM Technology Corp., Macedon, NY) was used to determine NDF and ADF for the calibration set (Vogel et al., 1999). Ash and ADL were determined for the calibration set as described in Van Soest et al. (1991) by placing the ANKOM bags containing the residual of the ADF procedure in a 3 L Daisy II incubator jar (ANKOM Technology Corp.) and covering them with 72% H₂SO₄. The samples were rotated in the incubator for 3 h, subsequently washed in hot water for 15 min followed by acetone for 10 min, dried in a 100°C oven overnight, and weighed after cooling to room temperature. Finally the entire sample bag with its remaining material was ashed at 525°C for 4 h, and the ash was weighed. Ash weights were calculated after accounting for the sample bag material. Acid detergent lignin was adjusted for ash. The standard error of calibration, $R^2$, and standard error of cross validation for the trait values predicted by NIRS are as follows: NDF = 0.34, 0.98, and 0.86; ADF = 0.22, 0.99, and 0.73; ADL = 0.15, 0.97, and 0.24.

**Genotyping**

We used simple sequence repeat (SSR) markers for genotyping using previously described markers (Julier et al., 2003; Robins et al., 2007b; Sledge et al., 2005) and additional markers designed from ESTs by ourselves and the Samuel Roberts Noble Foundation (M.J. Monteros, personal communication, 2009). Primers for all markers (Supplemental Table S1) were synthesized by Integrated DNA Technologies (Coralville, IA; http://www.idtdna.com/Home/Home.aspx [verified 6 Jan. 2011]) with the addition of 18 nucleotides of the M13 universal primer sequence (Schuelke, 2000) onto the 5′ end of the forward primer. The M13 universal primer sequence was synthesized and labeled with a blue (6-FAM), green (HEX), or yellow (NED) fluorescent tag by Applied Biosystems (Foster City, CA; http://www.appliedbiosystems.com [verified 30 Dec. 2010]). Polymerase chain reaction (PCR) recipes and ingredients were exactly the same as Sledge et al. (2005) and PCR programming was either the same as Julier et al. (2003) (4 min at 94°C, followed by 35 cycles with 30 sec at 94°C, 1 min at 55°C, and 1 min at 72°C, plus a final elongation step of 7 min at 72°C [program A in Supplemental Table S1]) or as follows: 95°C for 2 min, followed by 30 cycles with 30 sec at 95°C, 45 sec at 60°C, and 45 sec at 72°C, plus 10 cycles with 30 sec at 95°C, 45 sec at 53°C, 45 sec at 72°C, and finalized with an elongation step of 7 min at 72°C (program B in Supplemental Table S1). Each SSR marker was amplified by PCR independently, and we pooled PCR products of 4 to 6 reactions for genotyping on an automated ABI3730 sequencer at the University of Georgia (Athens, GA) DNA Sequencing Facility. The data files from the analyzer were analyzed by Genemarker software (SoftGenetics, State College, PA; www.softgenetics.com [verified 30 Dec. 2010]). Each allele of each marker was scored as “1” for presence and “0” for absence on each individual in the population.

**Data Analysis**

**Phenotype**

Total yearly biomass yield was measured in five environments: IA05, IA06, NY05, NY06, and Que06. The yearly yield for all environments was fit to a generalized linear model:

$$y_{ijk} = \mu + I_j + r_{(ijk)} + g_i + g_l_{ij} + e_{ijk},$$

in which $y_{ijk}$ is the yield for the $i$th genotype in the $k$th replication of the $j$th environment, $\mu$ is the grand mean, $I_j$ is the effect of the $j$th environment, $r_{(ijk)}$ is the effect of kth replication at the $j$th environment, $g_i$ is the genetic effect of the $i$th genotype, $g_l_{ij}$ is the interaction effect of $i$th genotype and $j$th environment, and $e_{ijk}$ is the residual. All factors were considered as random, and the ANOVA was performed using PROC GLM (SAS Institute, 2004). The variance components of residual ($\sigma^2_e$), genotype × environment interaction ($\sigma^2_{g_l}$), and genotype ($\sigma^2_g$) were estimated and used to compute broad-sense heritability as:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_e + \sigma^2_{g_l} + \sigma^2_g}.$$

Because significant genotype × environment interaction was observed in model [1], the yearly yield at each environment was fit separately to a simple generalized linear model:

$$y_{ik} = \mu + r_k + g_i + e_{ik},$$

in which $y_{ik}$ is the yield for the $i$th genotype in the $k$th replication, $\mu$ is the grand mean of that environment, $r_k$ is the effect of the $k$th replication, $g_i$ is the genetic effect of the $i$th genotype, and $e_{ik}$ is the residual. The genotype effect $g_i$ was considered as fixed. The least square mean of yearly yield for each of 190 genotypes, $M_i$, was estimated in each environment using PROC GLM (SAS Institute, 2004) and was used for association analysis.

Because the cell wall composition was measured only in one environment, the data were fit to the simple model [2]. The least square mean of each genotype was estimated using PROC GLM (SAS Institute, 2004) and used for association analysis. To estimate heritability, the genotype effect $g_i$ was considered as random, and the variance components of residual ($\sigma^2_e$) and genotype ($\sigma^2_g$) were estimated. The broad sense heritability was calculated as:
Marker Polymorphism

The number of alleles produced by each of the 71 SSR markers was counted, and frequency of each allele was computed across the entire population. Alleles with a frequency less than 10% (19 of 190 individuals) were considered to be rare alleles and were excluded from further analysis.

Linkage Disequilibrium Analysis

Because of complex genetic segregation in autotetraploid alfalfa, the genotypic LD between pairs of SSR markers was estimated using Fisher’s exact test. The test was based on the observed pairwise genotype frequency. We could not determine allele copy number, so the genotypic classes were actually phenotypes. For example, the marker phenotype A1A1A1A2 could include the following genotypes: A1A1A1A2, A1A1A2A1, A1A2A1A1, and A1A2A2A1. Genotypic LD was evaluated between pairs of SSR markers. To estimate physical locations of SSR markers, primer sequences were used to search *Medicago truncatula* Gaertn. pseudomolecules produced from the *M. truncatula* euchromatic genome sequence, build 3.0 with BLAST at the website (http://medicago.org/genome/cvU_blast.php [verified 30 Dec. 2010]). If both forward and reverse primers hit the same bacterial artificial chromosome (BAC) location with a predicted amplicon size similar to our observed fragment size, we used the position of the primer with the smaller number on the alignment with the BAC sequence as the physical location of the corresponding marker (Supplemental Table S1).

Population Structure and Kinship

Based on the LD between pairs of SSR markers, we picked 56 independent markers for population structure and kinship estimation. Population structure was calculated by STRUCTURE 2.2 (Falush et al., 2007; Pritchard et al., 2000). The number of subpopulations (z) was set from 1 to 20, and each subpopulation number (z) was evaluated for 20 repetitions using the admixture model and correlated allele frequencies. The iteration number for the Markov Chain Monte Carlo algorithm was set as 100,000, following a burn-in period of 100,000 iterations. Twenty more repetitions were performed for the z from 1 to 10 with the same parameters setting above. The log likelihood value and an ad hoc quantity (\(\Delta K\)) based on the second order change of log likelihood (Evanno et al., 2005) were calculated to estimate the number of subpopulations. After determining the number of subpopulations, the first \(z - 1\) columns of the population structure (S) matrix was used as the population structure (S) matrix and taken account into the linear mixed models for association analysis between markers and traits. The kinship coefficient matrix (A) was calculated by SPAGeDi (Hardy and Vekemans, 2002). Kinship coefficients lower than zero were set to zero.

Association Mapping

Because interactions between genotype and environment were observed for yield, we estimated least square means for yield in each environment using linear mixed models (Table 1) to evaluate associations. Model E1, the simplest model to explain yield variation with marker alleles, does not include either population structure (Q) or kinship (K) among the genotypes. Model E2 includes the kinship matrix, Model E3 includes the population structure matrix, and Model E4 includes both population structure and kinship matrices. For all models, the allele substitution effect and population structure were considered to be fixed effects; the genetic effect was considered random. The variance for the random effects for models E2 and E4 were assumed to be \(\text{Var}(\hat{g}) = 2\sigma^2_{g}\) and \(\text{Var}(\hat{\varepsilon}) = \text{I}\sigma^2_{\varepsilon}\), in which A is a 190 x 190 matrix of estimated kinship coefficients that defined the degree of genetic covariance between pairs of genotypes and I is a 190 x 190 matrix in which the off-diagonal elements were zero and the diagonal elements were one. The variances for the random effects for models E1 and E3 were assumed to be \(\text{Var}(\hat{g}) = \text{I}\sigma^2_{g}\) and \(\text{Var}(\hat{\varepsilon}) = \text{I}\sigma^2_{\varepsilon}\), respectively. Cell wall composition (NDF, ADF, and ADL) was measured at one environment, so the model for association is same as that for yield at each environment. Alleles associated with each trait were identified at \(p < 0.005\).

We programmed a macro in SAS to perform association analysis between alleles from the 71 SSR markers and each trait for all four mixed model equations with PROC MIXED (SAS Institute, 2004). The autotetraploid, noninbred alfalfa genotypes evaluated in this experiment could have more than two alleles at one SSR marker in a given individual. The dosage of any specific allele in the population could range from 0 to 4 for any one individual and could not be clearly determined. Thus, proper statistical analysis is difficult to perform and precisely identifying the alleles associated with target traits is problematic. In this study, we fitted trait data to a mixed model for each marker, which included all alleles for that marker (Pajerska-Mukhtar et al., 2009; Stich and Melchinger, 2009). This enabled us to test the effect of each allele of a given marker while removing the effects of other alleles from the same marker, providing a better association analysis compared to considering each allele as a separate locus. We evaluated the Bayesian information criterion (BIC) to compare the goodness of fit of the four different models.

Results

Phenotypic Evaluations

A significant genotype x environment interaction was observed for total yearly yield (Table 2). As a consequence, we considered the five location–year combinations as separate environments in the subsequent analyses. The 190 genotypes differed for all traits (\(p < 0.001\); Table 2). The broad-sense heritabilities were high for all traits: 86% for yield, 77% for NDF, 75% for ADF, and 74% for ADL (Table 2). All traits were normally distributed (Fig. 1). The range of least square means among 190 genotypes was great for all traits, spanning more than an order of magnitude for yield.
Heterozygosity and its Correlation with Phenotypes

The average allele number per genotype across the 71 SSR markers (heterozygosity) varied from 1.68 to 2.48, with an average of 2.10. The correlation between heterozygosity and yield was positive for all five environments, ranging from $r = 0.14$ to 0.20. However, no correlation was observed between heterozygosity and NDF, ADF, and ADL (Fig. 2).

Simple Sequence Repeat Genotyping and Marker Polymorphism

In total 470 alleles were scored across 71 SSR markers. Allele number per marker varied from 2 to 16, with 6.62
alleles per marker locus on average. Rare alleles present in fewer than 10% of individuals represented 33.6% of all alleles. Considering only the remaining 312 alleles, individual SSR markers produced between 2 and 8 alleles per marker, with an average of 4.39.

**Linkage Disequilibrium**
The frequency of marker pairs for which LD was observed using Fisher’s exact test decreased as the physical distance increased, as expected (Fig. 3). At a probability level of less than 0.001, 61.5% of marker pairs separated by less than 1 Mbp were in LD, with very few marker pairs showing LD for larger physical distances (Fig. 3).

**Population Structure and Kinship**
Based on 56 independent SSR markers, most (96.3%) pairwise kinship estimates between genotypes were smaller than 0.05, 3.0% were between 0.05 and 0.1, and only 0.7%
were between 0.1 and 0.35. We examined the possibility that the overall population of 190 individuals was structured into subpopulations ($z$) using STRUCTURE. As $z$ increased, the log likelihood value of each number of subpopulations being the true value increased smoothly (Fig. 4a), which suggested no population structure. To further investigate population subdivision, we calculated $\Delta K$ (Evanno et al., 2005), which indicated that the population could be partitioned into two subpopulations (Fig. 4b).

**Association Mapping**

We performed association analysis between each of the 312 nonrare alleles and each phenotypic trait. Based on the goodness of fit among the four mixed model equations, model E4 was the best for all traits, given that it had the smallest BIC value (Table 3). In contrast, model E1 was generally the worst, suggesting that accounting for both population structure and kinship was desirable, even though both were relatively small effects.

More alleles were associated ($p < 0.005$) with trait phenotypes in model E1 than in model E4 for all traits, which indicated that some alleles identified in model E1 could have been associated with traits due to population structure and/or genotypic relatedness rather than due to true genetic variation in the trait (Table 3). Model E2 showed greater improvement in terms of BIC and the number of significant alleles relative to model E1 than did model E3 for most traits, suggesting that kinship was more effective to control than population substructure in this population (Table 3). Model E4 showed slight improvement for some traits compared to model E2 in terms of BIC and the number of significant alleles (Table 3).

Based on model E4, 15 alleles derived from 13 markers showed strong association ($p < 0.005$) with yield in at least one of five environments (Fig. 5). Twelve markers could be located in the genome based on their physical location on the *M. truncatula* genome and/or their position on genetic linkage maps of diploid alfalfa (X.)

---

**Figure 3.** The frequency of significant tests for genotypic linkage disequilibrium (LD) between pairs of simple sequence repeat (SSR) markers separated by different physical distances across a range of $p$-values for an alfalfa breeding population.

**Figure 4.** Population structure analysis of an alfalfa breeding population. The top panel shows the average log–likelihood value and the bottom panel shows the ad hoc quantity, $\Delta K$, for differing numbers of subpopulations ($z$) within the population.
Li and E. C. Brummer, unpublished data, 2010) (Fig. 5). Most alleles showed associations across multiple environments (Fig. 5). Allele aw688546_301 showed strong association with yield in four environments (2 yr each in IA and NY) and a weak association ($p < 0.05$) in the fifth environment (Que06). Another allele from the same marker, aw688546_304, showed strong association with yield of NY05 (Fig. 5). Two alleles from marker bg281 (205 and 206) both showed association with yield in some New York and Québec environments (Fig. 5). Allele bg455405_159 showed strong association at New York, moderate association ($p < 0.01$) at Québec, and weak association at Iowa (Fig. 5). Alleles aw694047_213 and aw690665_232 showed strong association with yield both years in Iowa and a weak association in Québec (Fig. 5).

Based on model E4, no allele was associated with NDF at $p < 0.005$ (Fig. 5). Only one allele, al375136_169, showed strong association with ADF (Fig. 5). This allele was also associated with NDF at $p < 0.05$ (Fig. 5). For both traits, the allele al375136_169 had a positive effect, meaning that the presence of the allele increased NDF and ADF. Only one allele, be318471_266, showed association with ADL at $p < 0.005$ (Fig. 5). Interestingly, marker be318471 showed strong association with both yield and ADL. Individuals carrying allele be318471_266 tended to have lower ADL concentrations, while individuals

<table>
<thead>
<tr>
<th>Model</th>
<th>IA05</th>
<th>IA06</th>
<th>NY05</th>
<th>NY06</th>
<th>Que06</th>
<th>NDF</th>
<th>ADF</th>
<th>ADL</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>2753.6 (9)</td>
<td>2764.5 (10)</td>
<td>2720.1 (8)</td>
<td>2749.2 (6)</td>
<td>2873.8 (11)</td>
<td>757.8 (1)</td>
<td>628.1 (2)</td>
<td>324.6 (3)</td>
</tr>
<tr>
<td>E2</td>
<td>2741.6 (5)</td>
<td>2758.5 (6)</td>
<td>2707.6 (4)</td>
<td>2674.4 (5)</td>
<td>2807.7 (7)</td>
<td>748.5 (0)</td>
<td>619.8 (1)</td>
<td>299.6 (1)</td>
</tr>
<tr>
<td>E3</td>
<td>2741.8 (8)</td>
<td>2751.7 (8)</td>
<td>2708.3 (7)</td>
<td>2736.5 (7)</td>
<td>2861.0 (13)</td>
<td>756.6 (1)</td>
<td>629.5 (2)</td>
<td>325.8 (3)</td>
</tr>
<tr>
<td>E4</td>
<td>2717.7 (4)</td>
<td>2743.3 (5)</td>
<td>2688.9 (4)</td>
<td>2666.1 (5)</td>
<td>2794.2 (5)</td>
<td>744.3 (0)</td>
<td>622.6 (1)</td>
<td>299.3 (1)</td>
</tr>
</tbody>
</table>

*IA, Iowa; NY, New York; Que, Québec; 05, 2005; 06, 2006.

Table 3. The Bayesian information criterion (BIC) for four models assessing marker-trait associations. In parentheses are the numbers of alleles that showed associations at $p < 0.005$ for each test.
carrying a different allele, be318471_257, tended to have higher yield in Iowa.

In general, individual markers explained 2 to 6% of the variation in yield in this population when evaluated in multiple regression models (results not shown). This information is of limited value given the relatively small population size and the likely overestimate of the effects as well as the unknown linkage distances between marker and trait. The results suggest, however, that biomass yield is likely controlled by many genes of small effect.

Discussion

Heterozygosity and Its Correlations with Phenotypes

The average allele numbers per genotype across the 71 SSR markers is 2.10 in our population, which is similar to the heterozygosity we previously found in three diverse tetraploid cultivated alfalfa genotypes, which had 1.92 to 2.15 alleles per marker per genotype (Li et al., 2009). In this study, we found that yield had small but positive correlations with heterozygosity in all five environments, but NDF, ADF, and ADL were not correlated with marker heterozygosity. Previously, a correlation between yield and heterozygosity, estimated by restriction fragment length polymorphism (RFLP) markers, was also found in an experimental tetraploid alfalfa population (Kidwell et al., 1994). Heterosis is often observed for biomass yield in alfalfa (e.g., Riday et al., 2002; Riday and Brummer, 2005) but not for composition (Riday et al., 2002). The correlation between yield and heterozygosity supports the idea that complementary favorable alleles in repulsion phase at linked loci contribute to yield and heterosis for yield (Bingham et al., 1994; Li and Brummer, 2009).

Linkage Disequilibrium

Linkage disequilibrium in plants has been shown to extend from hundreds of base pairs to hundreds of kilobase pairs, depending on the species and population type analyzed (Alm et al., 2003; Hyten et al., 2007; Ingvarsson, 2005; Liu and Burke, 2006; Mather et al., 2007; Morrell et al., 2005; Nordborg et al., 2002; Remington et al., 2001; Simko et al., 2006). Biological factors such as recombination rate, genetic drift, selection, mutation, and population admixture all affect LD (Flint-Garcia et al., 2003). Because autogamous species have a low effective recombination rate, they typically have more LD than allogamous species (Nordborg, 2000).

We found LD (p < 0.001) between 61.5% of SSR marker pairs separated by less than 1 Mbp. In this population, the extent of LD was potentially affected by selection and genetic drift. The synthetic tetraploid alfalfa population used in this study was derived from 300 individuals of three cultivars (100 individuals from each cultivar). Each of the three cultivars resulted from selection and a restricted genetic base that could have resulted in genetic drift from a broadly-based panmictic population. Further population restriction and unintentional selection leading to the development of the population under evaluation could have further increased the extent of LD observed in this study. Additionally, cultivated populations often have more LD than wild populations, as shown in sunflower (Helianthus annuus L.) (Liu and Burke, 2006) and barley (Hordeum vulgare L.) (Caldwell et al., 2006), which could account for the relatively long extent of LD in this population.

In maize, relatively high levels of genome-wide LD were seen using SSR markers, even though LD rapidly declines between single nucleotide polymorphisms (SNPs) within genes. This result may be explained by a higher mutation rate in SSRs than at SNP locations (Remington et al., 2001). Thus, another possible explanation for the length of LD in our population may be because of high mutation rates of SSR markers. Further, LD decay is expected to increase as the number of founders increases, as was seen with a synthetic ryegrass (Lolium perenne L.) population (Auzanneau et al., 2007). Further, the effects of bottlenecks and selection on LD are diminished in tetraploid alfalfa due to the presence of four homologous copies per chromosome per plant. These factors all suggest that the levels of LD we observed may not reflect the true extent of LD that might be seen based on DNA sequence.

The extent of LD will dictate whether a sufficient number of markers can be evaluated to scan the entire genome or if association analyses need to be restricted to a selected set of candidate genes. Rapidly declining LD will necessitate either a very large number of markers to adequately cover the whole genome or the use of SNP targeted to candidate genes, the latter of which will only assay a small part of the genome, albeit at loci that may be disproportionately associated with the trait. One estimate of LD in a CONSTANS-LIKE gene in alfalfa suggested that LD decayed within the length of the gene (Herrmann et al., 2010). That result is based on sequencing of 400 individuals derived from 10 different cultivars, perhaps a broader genetic pool than that used in our experiment.

Our estimate of genome size is biased. We know that the 1C genome size of alfalfa is about 0.85 pg (Blondon et al., 1994) or about 830 Mbp, and the 1C genome size of M. truncatula cv. Jemalong is 0.575 pg (Blondon et al., 1994) or about 560 Mbp. Thus, the alfalfa genome is about 150% the size of M. truncatula. We computed physical distances between markers based on the M. truncatula reference genome, build 3.0. The genetic maps of alfalfa and M. truncatula are highly colinear (Choi et al., 2004), so we can assume that distances between the markers in the alfalfa genome are roughly proportional to those in M. truncatula. If we further assume that LD extends up to 1 Mbp, then we will need ~1000 markers to fully saturate the genome for association mapping.

For routine use in breeding programs—for instance, as applied to genome-wide selection (Heffner et al., 2009)—this number of markers may not be cost effective. If our estimate of LD is overly optimistic, then we have little hope of using genome-wide markers in breeding populations like this one in the foreseeable future. How well this
population reflects other commercial breeding populations is not known, although the methods used to create this population are typical of those used by the alfalfa breeding community (e.g., Rumbaugh et al., 1988).

**Population Structure**

Our population does not show strong evidence of subpopulation structure because tests for subpopulation numbers (z) resulted in smoothly increasing log likelihood values as z increased. Although the ad hoc quantity (ΔK) suggested that subpopulation number might be two, adding a population structure effect to our model had little improvement in terms of the goodness of fit and number of alleles associated with traits. Thus, we can conclude that there was no clear population structure in our breeding population. Clear population structure was not present in collections of elite alfalfa, barley, and potato (*Solanum tuberosum L.*) breeding germplasm (Flajoulot et al., 2005; Kraakman et al., 2004; Malosetti et al., 2007), possibly because of the relatively narrow genetic base in those elite breeding germplasm (Malosetti et al., 2007). In contrast, clear population structure is generally observed in broad-based germplasm collections (Comadran et al., 2009; Remington et al., 2001). We observed very clear substructure in a broad-based collection of wild diploid alfalfa germplasm (Sakiroglu et al., 2010). In this study, our population was derived from the random mating of a set of founders originating from three elite tetraploid cultivars, all of which were adapted to the northern part of the alfalfa growing area in the United States. This relatively narrow genetic background of cultivated germplasm combined with two generations of random mating apparently removed any substructure within the population. A similar lack of substructure was identified among a collection of seven alfalfa cultivars from France (Flajoulot et al., 2005).

**Association Mapping**

The main challenge of association mapping is to separate associations between markers and traits that are real, resulting from physical linkage of genes (which is what we want), from those that are ephemeral as a result of population structure and/or kinship relatedness. To address this problem, we used the QK mixed model for association analysis to control for population structure and kinship (Yu et al., 2006). The QK method effectively eliminated spurious associations caused by population structure and kinship in previous experiments (Stich and Melchingner, 2009; Stich et al., 2008; Yu et al., 2006; Zhao et al., 2007). In this study, the model E4, which included both population structure and kinship, had better fit compared to other three models. Consistently, some alleles associated with traits that were identified in model E1 (no population structure and kinship), model E2 (with kinship only), or model E3 (with population structure only) did not show significant association in model E4, suggesting that some positive marker–trait associations were caused by population structure and/or kinship. In terms of goodness of fit and reducing false positive associations, model E2 showed great improvement relative to model E1, but model E3 did not show a similar improvement over E1. This indicated that kinship was the key source that caused spurious association rather than population structure in this population.

Yield data were collected at three locations and in 2 yr for a total of five environments. Several alleles were associated with yield across multiple environments based on model E4. Many alleles showed genotype × environment interaction because they were identified in one (or a few) location but not all. Several markers also showed association with multiple traits. In the case of marker be318471, one allele was positively associated with yield and another allele negatively associated with ADL concentration in the stem. This result suggests that different alleles could be selected for the same marker locus when performing selection for multiple traits. One particularly encouraging result was that marker aw688546 previously associated with yield in a biparental mapping population (Robins et al., 2007a) was also associated with yield in this experiment.

**Conclusions**

This alfalfa breeding population shows no clear population substructure, a relatively long extent of LD based on SSR markers, and abundant variation for yield and cell wall composition. Thus, a genome scan approach to identify markers linked to traits should be feasible in this population. Because this is a breeding alfalfa population, alleles associated with traits are already in the breeding program, and their frequency could be easily increased using marker-assisted selection. However, based on our estimated LD level, we would need about 1000 markers to explore the whole alfalfa genome for association between markers and traits. Of the 71 SSR markers we evaluated, however, several associations with yield and cell wall composition were identified. To fully identify markers strongly linked to QTL and efficiently improve the target traits by marker-assisted selection, more markers are needed.

**Supplemental Information Available**

Supplemental material is available free of charge at http://www.crops.org/publications/tpg.

**Acknowledgments**

We thank Mark Smith for field plot assistance and Trish Patrick for compositional analysis assistance.

This research was part of the NE100 Regional Hatch Project “Breeding and Genetics of Forage Crops to Improve Productivity, Quality, and Industrial Uses,” and was supported by USDA-DOE Plant Feedstock Genomics for Bioenergy award 2006-35300-17224 to ECB. Xuehui Li and Yanling Wei contributed equally to this work.

**References**


**The Plant Genome • March 2011 • Vol. 4, No. 1**


associated with field resistance to late blight in populations of tetraploid potato cultivars. Genetics 181:1115–1127.


