Single-Feature Polymorphism Discovery in the Transcriptome of Tetraploid Alfalfa

S. Samuel Yang, Wayne Wenzhong Xu,* Mesfin Tesfaye, JoAnn F. S. Lamb, Hans-Joachim G. Jung, Deborah A. Samac, Carroll P. Vance, and John W. Gronwald*

Abstract
Advances in alfalfa [Medicago sativa (L.) subsp. sativa] breeding, molecular genetics, and genomics have been slow because this crop is an allogamous autotetraploid (2n = 4x = 32) with complex polysomic inheritance and few genomic resources. Increasing cellulose and decreasing lignin in alfalfa stem cell walls would improve this crop as a cellulosic ethanol feedstock. We conducted genome-wide analysis of single-feature polymorphisms (SFPs) of two alfalfa genotypes (252, 1283) that differ in stem cell wall lignin and cellulose concentrations. SFP analysis was conducted using the Medicago GeneChip (Affymetrix, Santa Clara, CA) as a cross-species platform. Analysis of GeneChip expression data files of alfalfa stem internodes of genotypes 252 and 1283 at two growth stages (elongating, post-elongation) revealed 10,890 SFPs in 8230 probe sets. Validation analysis by polymerase chain reaction (PCR)-sequencing of a random sample of SFPs indicated a 17% false discovery rate. Functional classification and over-representation analysis showed that genes involved in photosynthesis, stress response and cell wall biosynthesis were highly enriched among SFP-harboring genes. The Medicago GeneChip is a suitable cross-species platform for detecting SFPs in tetraploid alfalfa.

Alfalfa [Medicago sativa (L.) subsp. sativa], a member of the Fabaceae, is the most widely cultivated forage legume in the world (Michaud et al., 1988) and the fourth most widely grown crop in the United States (Samac et al., 2006). In 2008, over 60 million Mg of alfalfa dry hay with a value of over $10 billion were harvested from over 8.5 million hectares (NASS, 2008). In addition to being a valuable forage crop for livestock, alfalfa has considerable potential as a sustainable, cellulosic feedstock for ethanol production (Samac et al., 2006). The strategy for development of alfalfa as a bioenergy feedstock involves separating stems and leaves. The leaves would be used as a protein supplement for livestock while the stems would be used to produce ethanol. The efficiency of ethanol production from cellulosic biomass is positively correlated with cellulose content but negatively correlated with lignin content (Chappell et al., 2007; Chen and Dixon, 2007). The value of alfalfa as a cellulosic feedstock would be enhanced by developing new alfalfa varieties that have increased cellulose and decreased lignin in stem cell walls.


Abbreviations: ES, elongating stem internodes; EST, expressed sequence tag; FDR, false discovery rate; MAS, marker-assisted selection; PES, post-elongation stem internodes; PCR, polymerase chain reaction; PM, perfect match; SAM, significance analysis of microarrays; SFP, single-feature polymorphism; SNP, single nucleotide polymorphism.
The breeding of alfalfa varieties for traits desirable for cellulosic ethanol feedstocks will be challenging. In general, progress in the breeding of this crop for complex traits is slow. Alfalfa is an allogamous (cross-pollinated) autotetraploid (2n = 4x = 32) species with highly complex segregation and inheritance patterns (Rumbaugh et al., 1988). The perennial nature of alfalfa requires evaluation of germplasm over several years. As a result, the standard protocol for alfalfa breeding has a selection cycle of three to four years. Marker-assisted selection (MAS) with robust genetic markers for traits of interest, such as modified cell wall composition, would reduce the selection cycle to one year and increase genetic gain per cycle. However, there are a limited number of portable molecular markers (e.g., simple sequence repeat) available for MAS in cultivated tetraploid alfalfa (Diwan et al., 1997; Brouwer and Osborn, 1999; Brummer et al., 2001; Julier et al., 2003; Maureira and Osborn, 2004; Sledge et al., 2005). SNPs have not yet been identified in the alfalfa genome. Furthermore, there has been little research regarding alfalfa genomics. Tetraploid alfalfa has a genome of 800 to 900 Mbp with a basic set of eight chromosomes (1x = 8). Currently, there is no publicly funded alfalfa genome-sequencing project. As of August 2009, there were only 12,791 expressed sequence tag (EST) sequences available in the public database (http://www.ncbi.nlm.nih.gov/; verified 30 Sept. 2009).

Affymetrix GeneChips (Santa Clara, CA), originally developed for gene expression analysis, can also be used for the detection of genomic polymorphisms (Hazan and Kay, 2003). Typically, each probe set in the Affymetrix GeneChip consists of 11 perfect match (PM) and 11 mismatch (MM) 25-mer probes. These probes cover a portion of the coding region for the corresponding target gene. If a probe targets a polymorphic locus within a coding region, its hybridization signal intensity will differ from the signal of the nonpolymorphic locus. Therefore, it is possible to identify a probe that targets a polymorphic locus by scanning the signal intensity pattern among the 11 PM probes in each probe set. The polymorphisms detected by single probes in microarrays are called single-feature polymorphisms (SFPs). SFP discovery has been reported for various plants including Arabidopsis thaliana (Borevitz et al., 2003; West et al., 2006; Borevitz et al., 2007), barley (Hordeum vulgare L.) (Cui et al., 2005; Rostoks et al., 2005; Luo et al., 2007; Walia et al., 2007; Xu et al. 2009), domestic rice (Oryza sativa L.), (Kumar et al., 2007) and cowpea [Vigna unguiculata (L.) Walp] (Das et al., 2008).

For species such as alfalfa with large, unsequenced genomes and few genomic resources, SFP discovery is a rapid and cost-effective approach for whole-genome scanning of putative polymorphic loci. A high-density oligonucleotide microarray for alfalfa is not yet available. However, the Medicago GeneChip has been developed for Medicago truncatula, the diploid model legume species. Alfalfa and M. truncatula share the same eight orthologous basic chromosome sets but with different ploidy levels; 2n = 4x = 32 for alfalfa and 2n = 2x = 16 for M. truncatula. The Medicago GeneChip contains 50,916 probe sets designed from M. truncatula sequences, 1896 probe sets designed from alfalfa ESTs and 8,305 probe sets designed from 8226 transcripts from the symbiotic organism Sinorhizobium meliloti (http://www.affymetrix.com; verified 30 Sept. 2009). The Medicago GeneChip is a suitable platform for studying gene expression in alfalfa (Tesfaye et al., 2006; Tesfaye et al., 2009) and therefore can be used for SFP detection.

In this study, we conducted SFP analysis of Medicago GeneChip expression data from stem internodes of two alfalfa genotypes that differ in lignin and cellulosic concentration. To assist in identifying polymorphic candidate genes involved in cell wall synthesis, we also functionally classified the SFP-harboring genes. Over 10,000 SFPs were discovered with a false discovery rate (FDR) of 17%. Functional classification of the SFP-harboring genes indicated polymorphisms in over 200 cell wall genes, including genes involved in cellulosic and lignin synthesis. The SFPs discovered in this study provide a list of candidate genes that can be developed as molecular markers for use in a breeding program to improve alfalfa as a biofuel feedstock.

**MATERIALS AND METHODS**

**Plant Materials**

Alfalfa clonal lines 252 and 1283 were selected from population UMN 3097 that was created by mixing seed from six commercial alfalfa cultivars (5312, Rushmore, Magnagraze, Wintergreen, Windstar, and WL 325HQ). These clones were initially selected from a plant nursery at the University of Minnesota Sand Plains Research Farm, Becker, MN where individual plants were harvested twice yearly for two years at the 25% bloom stage of development. Clone 1283 was identified as having consistently low concentrations of cell wall glucose (cellulose) and Klason lignin, whereas clone 252 was consistently high for both cell wall components. Two subsequent replicated field studies including these clones verified their divergent cell wall glucose and Klason lignin concentrations (Lamb and Jung, unpublished).

**Cell Wall Analysis of Plants Grown in the Greenhouse**

The alfalfa clonal lines 252 and 1283 were propagated from cuttings and grown in the greenhouse. The greenhouse experiments consisted of three replicates arranged in a randomized complete block design. For each replicate, there were eight plants of each clone in individual pots. Stem internode tissues were harvested at full bloom and plant material for analysis was composited within each replicate. On the basis of tissue pliability and coloration, the internode in transition from elongation to post-elongation cambial growth was identified. Stem coloration changes from light-green to dark-green as development advances from the elongating to post-
The internodes immediately above [elongating stem (ES) internodes] and below this transition internode [post- elongation stem (PES) internodes] were collected for RNA extraction. Examination of ES and PES internodes by light microscopy confirmed that ES internodes contained cells undergoing primary cell wall development while PES internodes exhibited cell types associated with secondary xylem development (data not shown). All stem internodes below the sampled post-elongation internode were dried at 60°C and ground to pass a 1-mm screen in a cyclone-type mill prior to cell wall analysis. Briefly, cell wall polysaccharides were quantified using the Uppsala total dietary fiber method (Theander et al., 1995). Starch-free, 80% ethanol-insoluble samples were acid hydrolyzed to monosaccharide subunits. Neutral sugars were acetylated and the alditol acetate derivatives were quantified by GC-FID. Total uronic acids were quantified colorimetrically by the tartaric acid as a reference standard. Cellulose was estimated as cell wall glucose. Klason lignin was determined gravimetrically as the ash-free residue remaining after hydrolysis of the cell wall polysaccharides. Total cell wall concentration was calculated as the sum of neutral sugar residues, total uronic acids, and Klason lignin. For both alfalfa genotypes, stem cell wall components were compared by one-way ANOVA (analysis of variance) using a randomized complete block design.

**RNA Extraction, Labeling, and GeneChip Hybridization**

Elongating and post-elongation stem internodes, as defined above, were collected from alfalfa genotypes 252 and 1283 grown in the greenhouse. The same set of greenhouse plants used for stem cell wall analysis (described above) was used for RNA extraction and the three replicates of each clone were processed separately. After harvest, stem internodes were immediately frozen in liquid nitrogen. Total RNA was extracted from each sample using Trizol reagent (Invitrogen, Carlsbad, CA) following the manufacturer’s instructions. Contaminating genomic DNA was removed from each sample using RQ1 DNase kit following the manufacturer’s recommendations (Promega Inc, Madison, WI). RNA labeling, hybridization, washing, and scanning were conducted in the Microarray Facility, Biomedical Genomics Center, University of Minnesota following standard Affymetrix procedures. Briefly, 10 μg of total RNA were used to synthesize cDNA using SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Carlsbad, CA). Biotinylated cRNA was generated via in vitro transcription using the Enzo BioArray High Yield RNA Transcript Labeling Kit (Enzo Life Sciences, Farmingdale, NY). Fifteen micrograms of labeled RNA were chemically fragmented and used for hybridization. Affymetrix GeneChip Sample Cleanup Module was used for cDNA, biotinylated cRNA purification and chemical fragmentation. The integrity of total RNA and fragmented biotin-labeled cRNA was verified with RNA600 Nano LabChip using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Three replicate GeneChips were hybridized for each genotype and each tissue type (2 genotypes × 2 tissue types × 3 replicates = 12 GeneChips).

**SFP Detection**

Single-feature polymorphisms (SFPs) were identified using a method based on probe affinity difference and affinity shape power (Xu et al., 2009; https://dbw10.msi.umn.edu:8443/sfp; verified 30 Sept. 2009). Briefly, PM affinities for each probe were calculated from the normalized signal intensity and the probe set expression index based on the affinity model (Li and Wong, 2001; Hubbell et al., 2002; Irizarry et al., 2003; Cui et al., 2005). For all probes, a matrix of probe affinity differences between genotypes was created. To detect significant probe affinity differences among all probes in the affinity matrix, the SAM method (Tusher et al., 2001) in the Bioconductor Siggenes package (http://bioconductor.org; verified 30 Sept. 2009) was applied using an FDR cutoff of 0.1. For each probe that showed a significant difference in affinity, a two-value direction vector was created for each genotype. If the affinity value of the adjacent probe was higher than the affinity of the potential SFP, a value of 1 was assigned to the appropriate index of the vector. The shape power is the sum of the absolute values of the two-vector subtraction. A minimum value of 1.0 was assigned to the shape power. The original weight is the SFP affinity difference (Ad) divided by 30th percentile (pct) of the affinity differences of all probes within the same probe set. To obtain the final SFP weight score (W), the weight is multiplied by the shape power (P) raised to the power of two \[ W = \log_2 \left( \frac{Ad}{30 \text{pct}} \right) P^2 \] (Xu et al. 2009).

**SFP Verification by PCR and Sequencing**

To test the reliability of the SFP detection method used in this study, 42 SFP probes (29 from *M. truncatula* probe sets and 13 from *M. sativa* probe sets) with different weight score values were randomly selected, polymerase chain reaction (PCR) amplified and sequenced. Briefly, the primers spanning the SFP probe sequences were designed on the basis of an expected amplicon size of 200 to 300 bp using Primer Express version 2.0 (Applied Biosystems, Foster City, CA). The consensus sequences for the SFP probes selected were obtained from www.Affymetrix.com. The PCR (35 cycles of 1 min at 94°C, 1 min at 55°C, and 2 min at 72°C, with 5 min of pre-denaturation at 94°C, and 5 min of final extension at 72°C) was performed in an MJ Research Mini Gradient Thermal Cycler (Bio-Rad Laboratories, Hercules, CA). The PCR contained 40 pmol of each oligonucleotide, 5 μl 10×Taq DNA polymerase buffer (GenScript Corp., Piscataway, NJ), 200 μM of each dNTP, 1.5 mM MgCl₂, and 2.5 units Taq DNA polymerase (GenScript Corp., Piscataway, NJ) in a 50 μl volume. The amplicons were run on a 1.0% agarose gel. Only those amplicons from each alfalfa
functional and over-representation analysis of SFP-harboring genes

MapMan software (http://gabi.rzpd.de/projects/MapMan; verified 30 Sept. 2009) was originally developed to visualize Arabidopsis microarray data in multiple metabolic pathways and provide an overview of cellular function and regulation (Thimm et al., 2004). To adapt MapMan to the Medicago GeneChip system, Medicago GeneChip probe set consensus sequences (http://www.affymetrix.com/index.affx) were compared with the Genecip probe set utilizing the information available from the Arabidopsis proteome using BLASTX with an E-value cutoff of $10^{-10}$. For the top Arabidopsis hits, each Medicago probe set was allocated to one of the 34 major MapMan BINs following by assignment to the appropriate subBIN. In addition, the Medicago probe sets for gene families involved in cell wall development were identified using a BLASTX search (E-value $< 10^{-10}$) against protein sequences listed at the Purdue University Cell Wall Genomics website (http://cellwall.genomics.purdue.edu/; verified 6 Oct. 2009) and the Cell Wall Navigator at the University of California, Riverside (Girke et al., 2004; http://bioweb.ucr.edu/Cellwall/index.pl; verified 30 Sept. 2009). Putative transcription factor families on the Medicago GeneChip were identified using a BLASTX search (E-value $< 10^{-10}$) against 1827 protein sequences consisting of 56 transcription factor families from the Arabidopsis transcription factor database (http://datf.cbi.pku.edu.cn/index.php; verified 30 Sept. 2009). After the first round of classification, additional MapMan bins were manually assigned if needed via text search for GO, KEGG, COG, and MIPS ontology annotation for each Medicago probe set utilizing the information available at the GeneBins database (Goffard and Weiller, 2006). Additionally, Medicago probe sets for putative nodulin (or nodulin-like) genes were identified using known nodulin protein sequences (E-value $< 10^{-10}$) in other species such as soybean [Glycine max (L.) Merr.]. The new subBIN 26.31, misc.nodulins, was assigned following Tellström et al., (2007). Otherwise the sequence was classified as "not assigned."

PageMan, a software tool for comparative analysis of gene ontology, was originally developed to display and annotate overview graphs for profiling experiments (Usadel et al., 2006). We adapted PageMan to perform functional class over-representation analysis of SFP-harboring genes in alfalfa. First, SFP-harboring probe sets were given a value “one” and the rest of the probe sets were given a value “zero” as a false expression value. The data were then loaded into PageMan to perform Fisher’s exact test. This test examines whether the number of genes in a given class is due to random chance. Fisher’s exact test with Bonferroni correction with a z-value cutoff of “one” was applied to detect functional classes over- or under-represented among SFP-harboring genes in each genotype. Adjusted p-values produced were then transformed into their respective z-values. The resulting values were then false color-coded using a color scale of −6 to 6. Blue and red indicate over- and under-representation of the corresponding class, respectively.

RESULTS AND DISCUSSION
Alfalfa Genotypes

The alfalfa genotypes used in this study, 252 and 1283, are clonal lines selected for divergent stem cell wall concentration and composition under field conditions (see Materials and Methods). These genotypic differences in stem cell walls were environmentally stable for plants grown in the field (data not shown) and in the greenhouse (Table 1). On a dry matter and cell wall basis, cellulose (defined as glucose) was 17.5 and 5.8% greater, respectively, in cell walls of genotype 252 compared to genotype 1283 (Table 1). Expressed on a dry matter and cell wall basis, Klason lignin was 19.4 and 7.6% greater, respectively, in cell walls of genotype 252 compared to 1283. The increase in cellulose and lignin in genotype 252 was associated with an 11% increase in total cell wall dry matter. The results in Table 1 also indicate that the greater cellulose and lignin in stem cell walls of genotype 252 is associated with an increase in xylose (xylans) and a reduction in the concentrations of pectic sugar residues

Table 1. Comparison of cell wall components in stems of genotypes 252 and 1283 on a dry matter and cell wall basis.

<table>
<thead>
<tr>
<th>Component</th>
<th>Genotype 252</th>
<th>Genotype 1283</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell wall†</td>
<td>687</td>
<td>619</td>
<td>8</td>
</tr>
<tr>
<td>Klason lignin*</td>
<td>117</td>
<td>98</td>
<td>3</td>
</tr>
<tr>
<td>Glucose**</td>
<td>302</td>
<td>257</td>
<td>3</td>
</tr>
<tr>
<td>Klassen lignin†</td>
<td>170</td>
<td>158</td>
<td>2</td>
</tr>
<tr>
<td>Glucose†</td>
<td>440</td>
<td>416</td>
<td>2</td>
</tr>
<tr>
<td>Xylose†</td>
<td>145</td>
<td>138</td>
<td>1</td>
</tr>
<tr>
<td>Arabinose**</td>
<td>34</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>Galactose†</td>
<td>29</td>
<td>32</td>
<td>0.4</td>
</tr>
<tr>
<td>Mannose</td>
<td>32</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Rhamnose†</td>
<td>11</td>
<td>12</td>
<td>0.3</td>
</tr>
<tr>
<td>Fucose†</td>
<td>2.9</td>
<td>3.2</td>
<td>0.04</td>
</tr>
<tr>
<td>Uronic acids**</td>
<td>136</td>
<td>163</td>
<td>1</td>
</tr>
</tbody>
</table>

†Alfalfa clones differ at $P < 0.05$. Values are means based on duplicate analysis of three biological replicates for each clone. Plants were grown in the greenhouse (see Materials and Methods for details). SEM = Standard error of mean.

*Alfalfa clones differ at $P < 0.01$.

†Alfalfa clones differ at $P < 0.10$. 

genotype that exhibited a single band were further purified using the QuickClean 5M Gel Extraction Kit (GeneScript Corp., Piscataway, NJ) and then sequenced using an ABI PRISM 3130xl Genetic Analyzer (Foster City, CA) (DNA Sequencing and Analysis Facility, University of Minnesota). The sequence variation in the target SFP probes between two alfalfa genotypes was verified by examining sequence chromatograms.
(uronic acids, arabinose, galactose, and rhamnose). A major factor contributing to the increase in lignin and cellulose in stems of genotype 252 may be greater deposition of secondary xylem. As alfalfa stem internodes mature, cell wall concentration increases due to cambial activity and the deposition of xylem tissue. The thick secondary walls of this xylem tissue are rich in cellulose, xylans, and lignin, but contain little if any pectin (Jung and Engels, 2002).

**SFP Discovery**

cRNA was prepared from elongating and post-elongation stem internodes of genotypes 252 and 1283. Primary cell wall synthesis occurs in elongating internodes. In post-elongation internodes, growth in stem diameter is the result of the development of secondary xylem. The GeneChip cel files of gene expression from three replicates of elongating or post-elongation stem internodes from genotype 252 were compared with the data from genotype 1283 stem internodes at the same developmental stage. Figure 1 shows the application of the SFP detection algorithm used for data analysis in this study (see Materials and Methods for details). SFP weight scores were computed using the shape power for each SFP probe and putative SFPs were selected using a cutoff weight score of 2.5 (Supplemental Table S1). For example, the Mtr.10830.1.S1_at probe set contained several potential SFPs based on affinity difference (Fig. 1b). However, after applying a SFP weight score cutoff value of 2.5, only probes #5 and #6 were selected as SFPs. About half of the SFPs discovered had a weight score higher than 5.0 and nearly 1000 SFPs had a weight score of 40 or above (Fig. 2, Supplemental Table S1).

The number of SFPs detected from ES and PES internodes was 7174 and 8084, respectively. The number of SFPs common to both ES and PES internodes was 4368. When data obtained from comparisons of both stem internode tissues for each genotype were combined, a total of 10,890 SFPs in 8230 probe sets were identified (Supplemental Table S1). About 75% of the SFP probe sets contained a single SFP while about 15% contained two SFPs. As many as six SFPs were detected per probe set.

GeneChip probe sets often contain tiling probes with overlapping target regions. To estimate possible double-counting of SFPs on those tiling probes, we searched for tiling probes among the 10,890 SFP probes identified. We detected a total of 1200 tiling SFP probes with 9 to 24 overlapping bases (data not shown). Thus, there may be some double-counts within the 10,890 SFPs but the number should be small compared to the total number of SFPs detected. Because sequence information is not available, we could not evaluate the extent to which tiling contributed to double-counts.

**SFP Confirmation**

Because SNPs have not been identified in alfalfa, evaluating the FDR for the putative SFPs by comparison with known SNPs was not possible. Instead, 42 SFPs with different weight scores (29 from *M. truncatula* probe sets and 13 from *M. sativa* probe sets) were randomly selected for SFP verification by PCR-sequencing. PCR primers spanning the SFP probe target regions were designed with an expected amplicon size of 200 to 300 bp (Supplemental Table S2). Sequence chromatograms of the PCR products amplified from genomic DNA of genotypes 252 and 1283 were examined for polymorphisms. Figure 3

![Figure 1. SFP analysis plots of probe sets Mtr.41488.1.S1_at (a) and Mtr.10830.1.S1_at (b). Three replicates of genotype 252 and genotype 1283 are indicated by red and black lines. Raw data panel shows log2 intensity of PM (perfect match) probes before normalization. X axis represents the 11 PM probes. The affinity panel is the log2 normalized intensity subtracted by the expression index of that probe set. The affinity difference panel indicates the probe affinity of genotype1283 subtracted from the probe affinity of genotype 252. The weight score panel represents the SFPs after a cutoff value of 2.5 was applied. SFP detection algorithm used is available at University of Minnesota Supercomputing Institute website; https://dbw10.msi.umn.edu:8443/sfp/.


shows one of the sequence chromatograms examined for a putative SFP for probe #6 of probe set Mt.41488.1.S1_at. As shown in the affinity plot (Fig. 1a), there is a significant decrease in affinity for probe #6 for genotype 252 compared to genotype 1283. The probe target sequence contains “A” (Fig. 3a), the sequence for genotype 252 contains a single “T” peak at the SFP target site (Fig. 3b), while the sequence for genotype 1283 contains a mixture of a dominant “A” peak plus a small “T” peak at the SFP target site (Fig. 3c).

Of the 42 putative SFPs examined, 35 (83%) exhibited polymorphisms while seven putative SFPs were false positives (FDR = 17%) (Supplemental Table S3). The 17% FDR measured in this study is relatively low compared to FDRs reported in other studies using cRNA for SFP detection (Rostoks et al., 2005; Cui et al., 2005; Das et al., 2008). In two studies using the barley Affymetrix GeneChip to detect SFPs in barley, the FDR was estimated to be 10 to 20% (Cui et al., 2005) and 40% (Rostoks et al., 2005). In our study, we used a cross-species platform (Medicago GeneChip) for SFP detection in tetraploid alfalfa. Of the total probe sets on the Medicago GeneChip, 1896 were from alfalfa and 50,916 were from the model legume M. truncatula. Our report is the second where a cross-species platform has been used to detect SFPs in the transcriptome of plants. Previously, the soybean Affymetrix GeneChip was used to detect SFPs in cowpea (Vigna unguiculata L. Walp.) (Das et al., 2008). These authors estimated the FDR to be 32%, which is substantially higher than the 17% FDR found in this study. The lower FDR (17%) estimated by PCR sequencing validates the SFP detection parameter (weight score cut off 2.5) used in this study. The lack of alfalfa sequence information for primer design may also have contributed to the FDR detected in this study. The false positives may be due to amplification of duplicated genes (paralogs) in alfalfa genotypes 252 and 1283 that do not exhibit polymorphism. Alternatively, variation in post-transcriptional modification such as alternative splicing may have contributed to the FDR (Rostoks et al., 2005; Das et al., 2008).

**Functional Classification**

The SFPs discovered in this study significantly increases the number of polymorphisms that can be developed into molecular markers for MAS in alfalfa. These molecular markers can be used in candidate gene-based association mapping to identify markers useful for selecting alfalfa germplasm with modified stem cell wall composition. Candidate gene-based association mapping is hypothesis-driven and requires knowledge of the key genes regulating traits of interest (Long and Langley, 1999; Pfieger et al., 2001; Tabor et al., 2002; Neale and Savolainen, 2004; González-Martínez et al., 2007). In our case, the traits of interest are cell wall lignin and cellulose concentrations in alfalfa stems. Metabolic pathways involved in lignin and cellulose synthesis have been well characterized and many of the genes involved have been cloned in some species (Boerjan et al., 2003; Somerville, 2006).
To assist in identifying candidate genes with polymorphisms that could be developed into molecular markers in alfalfa, we performed functional classification of the SFP-harboring probe sets (see Materials and Methods for details). Numerous SFPs were identified in different gene functional classes (Fig. 4, Supplemental Table S4). SFPs were identified in more than two hundred cell wall genes including those involved in cell wall structure, cell wall modification, cellulose biosynthesis, cell wall precursor synthesis, and cell wall degradation. For example, cellulose synthase genes Mtr.5274.1.S1_s_at and Mtr.51462.1.S1_s_at contained four and three SFPs, respectively (Fig. 4, Supplemental Table S4). SFPs were also identified for genes involved in the synthesis of UDP-glucose (SuSy, sucrose synthase) and its conversion to UDP-glucuronic acid (UGD, UDP-glucose dehydrogenase) (Supplemental Table S4, Supplemental Figure S2). Both SuSy and UGD play key roles in cell wall synthesis.

Over-Representation Analysis

We adapted PageMan (Usadel et al., 2006), a software tool for comparative analysis of gene ontology, to perform functional class over-representation analysis of SFP-harboring genes in genotypes 252 and 1283. Fisher’s exact test with conservative Bonferroni correction (z-value cut off = 1) was applied to detect functional classes over- or under-represented among the SFP-harboring genes, and also for SFP-harboring genes that contained three or more SFPs (Fig. 5). The classes over-represented among SFP-harboring genes included cell wall, stress response, and photosynthesis. Functional classes of SFP-harboring genes containing at least three SFPs per probe set (hypervariable
genes) included miscellaneous and DNA synthesis/chromatin structure.histone classes.

CONCLUSIONS
SFP detection using the Medicago GeneChip is a rapid and cost-effective approach for genome-wide polymorphism discovery in tetraploid alfalfa, a species with a large unsequenced genome and few genomic resources. Using the Medicago GeneChip as a cross-species platform for alfalfa, we identified over 10,000 SFPs between two alfalfa genotypes with an estimated FDR of 17%. Functional classification of the SFP-harboring genes indicated that over 200 were involved in cell wall synthesis including those involved in cellulose and lignin synthesis. These SFPs provide a pool of candidate genes that can be developed as molecular markers (e.g. SNPs) for use in a breeding program to improve alfalfa as a biofuel feedstock.

Supplemental Tables
Supplemental Tables S1 to S4 are available at the University of Minnesota Supercomputing Institute website https://dbw10.msi.umn.edu:8443/sfp/Alfalfa.jsp.

Supplemental Figures
Supplemental Figures S1 and S2 are available at the University of Minnesota Supercomputing Institute website https://dbw10.msi.umn.edu:8443/sfp/Alfalfa.jsp.

Acknowledgments
We thank Dr. B. Usadel for providing the MapMan Mapping source file for Arabidopsis thaliana. This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute for Advanced Computational Research. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

References

Figure 5. Over-representation analysis of functional classes among SFP-harboring genes using the PageMan over-representation analysis module. Functional classes that are over- or under-represented among SFP-harboring genes are indicated. “Hypervariable genes” indicate over- or under-representation of genes that harbor at least three SFPs. Briefly, adjusted p-values produced by Fisher’s exact test with Bonferroni correction (cutoff value = 1.0) were transformed into z-values. The z-values for the functional class over- or under-represented in each group are provided in the parenthesis. The resulting values were then false color coded using a scale of –6 to +6. Blue and red indicate over- and under-representation of the corresponding class, respectively. See Materials and Methods for details.


