POPULATION STRUCTURE, GENETIC DIVERSITY, PHYLOGENETIC ANALYSES, AND ASSOCIATION MAPPING OF BIOFUEL TRAITS IN WILD DIPLOID ALFALFA (*MEDICAGO SATIVA* L.) ACCESSIONS

by

MUHAMMET ŞAKİROĞLU

(Under the Direction of E. Charles Brummer)

ABSTRACT

Cultivated alfalfa derives from a taxonomic group called the *Medicago sativa-falcata* complex. The complex consists of several species and subspecies that do not have any hybridization barriers. Morphological traits such as flower color, pod shape, and ploidy have traditionally being used for taxonomic classification. Cultivated alfalfa is tetraploid, but a significant amount of diversity is present among diploid germplasm. A collection of 374 individual genotypes derived from 120 unimproved diploid accessions from the National Plant Germplasm System was selected to represent the diploid *M. sativa-falcata* complex, including *M. sativa* subspecies *caerulea*, *falcata*, and *hemicycla*. The accessions were screened with a set of 89 polymorphic SSR loci in order to estimate genetic diversity, infer the genetic bases of current morphology-based taxonomy, and determine population structure. High levels of variation were detected. A model-based clustering analysis of the genomic data identified the morphologically defined subspecies falcata and caerulea. The hybrid nature of subspecies hemicycla has also been confirmed based on its genome composition. Subsequent hierarchical population structures indicated that two distinct subpopulations exist within subspecies caerulea

and subspecies falcata. We also evaluated performance of selected genotypes for cell wall constituents, total biomass yield, and other related agronomic traits and found a high amount of genetic variation in the diploid gene pool for agronomic traits and also for cell wall constituents. Large variation was present in this material, exceeding that observed in the tetraploid alfalfa core collection. Understanding patterns of linkage disequilibrium (LD) decay in alfalfa is necessary to determine the ability of association mapping to identify quantitative trait loci of important agronomic traits. We used SSR markers and sequence polymorphism in a lignin biosynthesis gene (F5H) to infer genomewide and within gene estimates of LD. We found extensive LD among SSR markers, extending over 10 Mb. In contrast, within gene LD extends over about 200 bp and sharply declined for longer distances. These results indicate that either more markers or more candidate genes are necessary in order to identify effective marker-trait associations.

INDEX WORDS: Alfalfa, Genetic Diversity, Population Structure, SSR, Diploid, Biofuel, Linkage Disequilibrium, Association Mapping

POPULATION STRUCTURE, GENETIC DIVERSITY, PHYLOGENETIC ANALYSES, AND ASSOCIATION MAPPING OF BIOFUEL TRAITS IN WILD DIPLOID ALFALFA (*MEDICAGO SATIVA* L.) ACCESSIONS

by

MUHAMMET ŞAKİROĞLU

BS, Harran University, Turkey, 2000

MS, Iowa State University, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2009

POPULATION STRUCTURE, GENETIC DIVERSITY, PHYLOGENETIC ANALYSES, AND ASSOCIATION MAPPING OF BIOFUEL TRAITS IN WILD DIPLOID ALFALFA (*MEDICAGO SATIVA* L.) ACCESSIONS

by

MUHAMMET ŞAKİROĞLU

Major Professor: E Charles Brummer

Committee: H. Roger Boerma
Steven J. Knapp
Katrien J. Devos
John Burke

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May, 2009

DEDICATION

This dissertation is dedicated to the memory of Seyyid Ahmet Şakiroğlu.

(Seyyid Ahmet Şakiroğlu'nun hatırasına)

ACKNOWLEDGEMENTS

I am grateful for the opportunity to work in the Brummer Lab. First of all, I wish to express my sincere gratitude to my major professor Dr. E. Charles Brummer, for his support throughout this work and also my entire graduate education.

I am grateful to the members who served on my committee and took the time to review this document including Dr. Roger Boerma, Dr. John Burke, Dr. Katrien Devos, and Dr. Steve Knapp. I would like to thank my colleagues Yanling Wei and Dr. Xuehui Li and other lab members. I would like to thank Donald Wood, Jonathan Markham, and Frank Newsome for the technical assistance at UGA and Trish Patrick at ISU. I will remain grateful to the Turkish Government for funding my graduate education, the National Research Initiative (NRI) Plant Feedstock Genomics for Bioenergy Program for funding this project.

I would like to give special thanks to my wife, Hülya, who has given me endless support and my sons, Mirza and Ahmet Bera, who never fail to bring joy to my life.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION

Alfalfa is one of the most important forage crops in the world with over 32 million hectares grown globally (Michaud et al. 1988). Cultivated alfalfa has been improved from a complex taxonomic group known as the *Medicago sativa-falcata* complex which includes several species and subspecies. Classification of taxa within the complex is based on the morphological traits of flower color, pod coiling, and pollen shape and on ploidy. The complex includes diploid and tetraploid taxa and interploidy hybridization is possible (McCoy and Bingham, 1988). Diploid subspecies (2n=2x=16) are *M. sativa* subsp. *falcata* (yellow flowers, sickle shaped pods), *M. sativa* subsp. *caerule*a (purple flowers, coiled pods), and their natural hybrid, *M. sativa* subsp. *hemicycla*. A related species, *M. glomerata* is also included in the group. The tetraploid subspecies (2n=4x=32) are *M. sativa* subsp. *sativa* (the direct analogue of diploid *caerulea*), *M. sativa* subsp. *falcata*, and the tetraploid hybrid *M. sativa* subsp. *varia*. The tetraploid version of *M. glomerata*, *M. glutinosa*, is included here (Quiros and Bauchan, 1988). The validity of current morphology-based classification of subspecies has not been confirmed with genomic tools.

In addition to having numerous forage qualities, alfalfa has recently been proposed as a bioenergy crop (Delong et al., 1995). Alfalfa stems and leaves can be mechanically separated with the leaves being used as a high protein animal feed and the stems to produce energy (McCaslin and Miller, 2007; Lamb et al., 2007). Alfalfa significantly reduces the need for fossil

fuel based synthetic nitrogen fertilizers, which can cause environmental problems, thereby decreasing the cost of production (Patzek, 2004; Crews et al., 2004). Understanding the synthesis of the plant cell wall components and the means to modify their quality and quantity toward a desired level will be vital for effective bioethanol production (Farrokhi et al., 2006).

We have selected 374 individuals from 120 accessions of unimproved diploid *M. sativa* accessions from throughout the Northern Hemisphere for evaluation. Our objective of the first study in this dissertation was to investigate the population structure of a wide range of diploid members of the *M. sativa-falcata* species complex and to test concordance between current morphology-based classification and differentiation based on SSR markers. We also intended to infer the extent of genetic diversity that exists in diploid accessions. Such a comprehensive study will help to determine if there is any genetic rationale underlying the current morphology-based taxonomy. It will also allow the evaluation of the diploid gene pool of cultivated alfalfa. Evaluation and understanding of population structure, allelic richness, and diversity parameters of diploid germplasm will help breeders to more effectively use genetic resources for cultivar development.

Our objective in the second experiment in this dissertation was to evaluate the performance of the selected individual genotypes in replicated trials in two locations over two years (i) to measure the boundaries of variation of biofeedstock traits of the diploid gene pool of cultivated alfalfa, (ii) to investigate the population structure of wild alfalfa based on phenotype, and (iii) to compare population structure based on phenotypic and genotypic data. Knowledge of the extent of variability will be very useful for selection of appropriate germplasm for subsequent genetic mapping purposes and for effective introgression of diploid germplasm into cultivated breeding pools.

Linking DNA polymorphism to the phenotypic variation in the traits of interest is invaluable for plant breeding programs (Lande and Thompson, 1990). Diploid alfalfa could be used to conduct genetic mapping to avoid the complicated inheritance patterns present in autotetraploid cultivated alfalfa. The genetic maps of diploid and tetraploid alfalfa are highly syntenic. Because hybridizing taxa with different ploidy levels is possible (Kaló et al., 2000), extrapolating knowledge of quantitative trait loci (QTL) from diploid to tetraploid cultivated alfalfa is possible. Association mapping takes advantage of many generations of historic recombination that decrease linkage disequilibrium (LD) to short chromosome intervals that can be very useful for creating strong and robust statistical marker-trait associations (Jannink and Walsh, 2002). Understanding the patterns of LD in the entire genome and within genes, and also within and among populations is critical to determine strategies for mapping underlying genes (Rafalski and Morgante, 2004). Unfortunately, the extent of LD in alfalfa is unknown.

As the third objective of this dissertation, we estimated the extent of LD among SSR loci distributed throughout the genome and in a ~500bp transcribed region of the *ferulate 5-hydroxylase (F5H)* gene that is known to be involved in lignin biosynthesis using the same set of 374 unimproved diploid alfalfa genotypes from 120 accessions. We used association mapping to detect SSR and SNP markers that were associated with 17 cell-wall and agronomic traits. We aim to understand patterns of LD both at the genomewide level and within gene level in order to test the applicability of association mapping in alfalfa and to identify possible marker-trait associations that can be integrated into future improvement programs of alfalfa for biofeedstock.

LITERATURE REVIEW

Alfalfa

Alfalfa, the oldest plant that has been exclusively grown for forage, is the most important forage legume in the world (Quiros and Bauchan, 1988; Michaud et al. 1988). Originating from a region that includes eastern Turkey, southern Caucasia, and northern Iran, it has long been used by older civilizations of the region such as the Egyptians, Medes, and Persians. At the time of invasion of Greece by Xerxes around 450 B.C., alfalfa was introduced into Europe. Introduction into the Americas was initiated by the Spanish into Mexico; however, cultivation in North America did not take place until mid-19th century (Coburn, 1903). Currently, alfalfa is the fourth most widely grown crop in the USA, following corn, soybean, and wheat, and its economic value exceeds $10 billion a year. It is mainly used as feed for farm animals, especially dairy cows (Barnes and Sheafer, 1995).

Cultivated alfalfa is mostly a tetrasomic tetraploid (2n=4x=32) but a few diploid cultivars have been released. Diploid populations also exist in nature. Alfalfa is an outbreeding species, with populations being heterogeneous mixtures of highly heterozygous individuals**.** Alfalfa exhibits severe inbreeding depression when self-fertilized (Rumbaugh et al., 1988). A self-incompatibility mechanism has been described in alfalfa, but genotypes range widely in self fertility (Brink and Copper, 1938; Rowlands, 1964).

Cultivated alfalfa belongs to a complex taxonomic group known as the *Medicago sativa-falcata* complex. The complex includes several subspecies that can freely hybridize. The classification of taxa in the *Medicago sativa* complex as species or subspecies has been controversial (Sinskaya, 1961; Lesins and Lesins, 1979; Ivanov, 1988; and Quiros and Bauchan, 1988). Sinskaya (1961) denoted caerulea, hemicycla, falcata, and *sativa* as species and described a “circle of species” which includes all the taxa above and a number of other taxa. The division

of taxa in the “circle of species” was based on ploidy, assuming that taxa within the same ploidy level were more closely related. Lesins and Lesins (1979) classified falcata and *sativa* as different species of genus *Medicago*; however, hemicycla and caerulea were relegated to the subspecific level. They also noted that despite the obvious morphological distinctness of the two, there exist no hybridization obstacles between falcata and *sativa*. More recently, the previously defined species have been given subspecies status within the *M. sativa-falcata* complex (Quiros and Bauchan, 1988).

The criteria for classification of taxa included in the complex, in addition to ploidy, are their morphological traits; primarily flower color, pod shape, and pollen shape. Both ploidy levels are present in the complex and interploidy gene flow by unreduced gametes is thought to occur (McCoy and Bingham, 1988). Diploid subspecies (2n=2x=16) are *M. sativa* subsp. *falcata* (yellow flowers, sickle shaped pods), *M. sativa* subsp. *caerule*a (purple flowers, coiled pods), and their natural hybrid, *M. sativa* subsp. *hemicycla*. A related species, *M. glomerata* is also included in the diploid group. The tetraploid subspecies (2n=4x=32) are *M. sativa* subsp. *sativa* (the direct analogue of diploid caerulea), *M. sativa* subsp. *falcata*, and the tetraploid hybrid *M. sativa* subsp. *varia*. The tetraploid version of *M. glomerata*, *M. glutinosa*, also associated with complex (Quiros and Bauchan, 1988; Stanfrord et al., 1972). The genetic maps of diploid and tetraploid alfalfa are highly syntenic, and together with the possibility of interploidy hybridization, extrapolation of genetic studies conducted on diploid alfalfa to cultivated tetraploid alfalfa should be possible (Kaló et al., 2000).

Genetic Diversity

Genetic diversity plays a key role in plant breeding. Absence of enough genetic diversity can significantly reduce the effectiveness of plant breeding for further crop improvement

(Hoisington et al., 1999). Despite the fact that the importance of genetic resources is widely appreciated and considerable efforts have been devoted to collection and maintenance of genetic variation, in fact, genetic resources have not been effectively used to improve yield and other complex traits (Tanksley and McCouch, 1997).

The exploration of genetic diversity and population structure in alfalfa has generally focused on tetraploid breeding populations or progenitor germplasm. Barnes et al. (1977) defined nine historical germplasm sources as the early introduction of alfalfa germplasm in the North America. The nine progenitor germplasms are *M. sativa* subsp. *falcata, M. sativa* subsp. *sativa*, Ladak, Flemish, Turkistan, Indian, African, Chilean, Peruvian, and *M. sativa* subsp. *varia* (Barnes et al., 1977). *Medicago falcata*, *M. varia*, Ladak, and Turkistan are considered to have contributed to increased fall dormancy of commercial cultivars, whereas, the Indian, African, Chilean, and Peruvian germplasm sources have contributed to nondormant cultivars (Barnes et al., 1977).

The genetic relationships among the nine germplasm sources show that subsp. *falcata* is genetically the most dissimilar to the others, but Peruvian germplasm is also somewhat unique (Kidwell et al 1994; Segovia-Lerma et al, 2003). The diversity analyses of four of the nondormant germplasms, Indian, African, Chilean, and Peruvian via comparative C-banding revealed no separation between nondormant germplasm except a weak separation of Indian from the rest of germplasm (Bauchan et al., 2003). Comparison of some contemporary cultivars with the historical introductions concluded that current U.S. cultivars have largely diverged from historical introductions over time (Mauriera et al 2004; Vandermark et al., 2006). Molecular markers have been used to differentiate Italian populations and ecotypes and to separate Italian and Egyptian cultivars (Pupilli et al., 1996; Pupilli et al., 2000). The within population genetic

variation of some Italian varieties and ecotypes evaluated with SSR markers was around 77% of total genetic variation (Mengoni et al. 2000). Eight SSR markers revealed low but highly significant values of F_{ST} between pairwise comparisons of very closely related tetraploid French cultivars (Flajoulot et al., 2005). Brummer et al. (1991) used 19 cDNA RFLPs to infer population structure of diploids and were able to differentiate *M. sativa* subsp. *falcata* from *M. sativa* subsp. *caerulea*.

Alfalfa Breeding

Alfalfa has the potential to produce high yield but genetic improvement for yield is not as high as those realized from the major grain crops (Hill et al 1988). Average yield increases per year in alfalfa have been reported to vary from <0.10 to 0.30% by different researchers (Hill et al. 1988, Holland and Bingham 1994, Volenec et al., 2002). However, Riday and Brummer (2002) and Lamb et al. (2006) suggested that due to pathogen and pest pressures, the yield of older cultivars were declining while the newer cultivars only maintained the yield level about the same. Lamb et al. (2006) also found that there were large location and year effects when testing cultivars released at different time periods in a multiple year and location experiment. They concluded that the only gains in yield occurred in those locations with significant disease pressure; without the biotic stress, yield was stagnant over 60 years. In fact, USDA data indicate no yield increase in alfalfa since 1983 (USDA, 2004). Hill et al. (1988) listed three possibilities contributing yield stagnation in alfalfa. The perennial growth habit of alfalfa could reduce yield increase rate, because evaluation of selection material must be done for several years before any selection can be done. Hence, gain per year in alfalfa will be significantly lower than that in an annual crop. A second possible explanation is that unlike grains, where only certain parts of the plant are being harvested, the total plant biomass is harvested in alfalfa. To increase yield in

grain crops, breeders can alter partitioning of yield between grain and the rest of the plant, rather than altering the assimilatory process as forage breeders must do. However, the third and most likely suggestion is that alfalfa breeders have focused on pest resistance and simply not selected for yield (Hill et al., 1988). Inbreeding depression is a limiting factor for developing inbred lines in alfalfa, and as a consequence, most alfalfa cultivars are synthetic cultivars (Brummer, 1999).

The most obvious way to increase yield is through population improvement via recurrent selection (Fehr, 1993), although limited selection for yield *per se* has been conducted in the past. A second method for yield increase is a semihybrid model, which proposes the evaluation of heterotic patterns between diverse alfalfa germplasm and the generation of population hybrids for sale (Brummer 1999). A considerable effort has been devoted to the identification of distinct germplasm that exhibit heterosis. Early heterosis studies focused on the heterotic potential between subspecies *sativa* and *falcata* and positive results have been reported (Westgate, 1910; Waldron, 1920; Sriwatanapongse and Wilsie 1968). More recently, Riday and Brummer (2002a) have found that falcata × sativa hybrids express forage yield heterosis, but due to undesirable characteristics of falcata germplasm, such as early dormancy and slow regrowth, usage of sativa × falcata hybrids in breeding is not currently feasible (Riday and Brummer 2002b). In order to overcome the limitations of falcata germplasm, Şakiroğlu and Brummer (2007) tested the potential of elite Midwestern cultivars to express heterosis when crossed to southwestern U.S. germplasm selected for adaptation to the Midwestern U.S., but unfortunately found no heterosis. A third possible means to overcome the current yield stagnation problem is the identification of QTL that affect biomass yield via QTL mapping and introgress those QTL into elite breeding pools via marker assisted selection.

There are a limited number of mapping studies in alfalfa to target marker trait association to date. Simply inherited traits such as unifoliate leaf (Brouwer and Osborn, 1997) and non-nodulating phenotypes (Endre et al., 2002) have been mapped. The RFLP loci linked to gene *uni* that causes the unifoliate leaf mutation were identified using bulked segregant analysis of an F_2 diploid alfalfa population (Brouwer and Osborn, 1997). There were a number of attempts to map QTL responsible for complex traits. Brouwer et al. (2000) were able to locate several QTL for winter injury, freezing injury, and fall growth in a backcross population of tetraploid alfalfa using Single Dose Restriction Fragments (SDRF). Sledge et al. (2002) identified four unlinked RFLP markers associated with aluminum tolerance in diploid alfalfa and two of those have been confirmed in backcross populations. Comparing 36 individual genotypes from a resistant population and another set of 36 nonresistant individual from another population, Obert et al. (2000) were able to detect association between four AFLP markers and mildew resistance in tetraploid alfalfa. Finally, Robins et al. (2007a, 2007b) have conducted QTL mapping in tetraploid alfalfa using parents from *M. sativa* subsp. *falcata* and *M. sativa* subsp. *sativa* using RFLP and SSR markers. Significant marker – QTL associations were detected for yield and for agronomic traits. Both parents contributed favorable alleles from complementary loci (Robins et al., 2007a, 2007b). Thus, identification of QTLs controlling desirable traits is possible via molecular markers and could contribute to breeding efforts.

Alfalfa as a Bioenergy Crop

Growing energy demand and environmental problems caused by extensive fossil fuel consumption have led researchers and administrators to seek economically and ecologically sustainable renewable energy sources. A number of renewable energy sources have been used in the past, including wind, solar, geothermal, hydropower, and biomass. Bioenergy is defined as

the usage of plants or plant derived materials to produce energy (Brown, 2003). It is the second largest renewable energy source in the world after hydropower, and annual biomass energy production is estimated to be 50 EJ per annum, or about 12% of the world's primary energy consumption of 406 EJ (U.S. Department of Energy: http://www.energy.gov/). Biomass can be used in two major ways to produce energy, by fermentation to produce liquid fuel ethanol or by direct burning to produce electricity or syngas (Brown, 2003). In either case, for plant biomass to be an economically viable option, yields must be maximized and the cell wall composition modified to increase the energy potential of plant material (Ragaukas et al., 2006).

Bioenergy crop production should be both economically profitable for growers and environmentally sustainable (Hanegraaf et al., 1998). Biomass used for energy will be a low value commodity, so simultaneous production of a high-value co-product is essential. For alfalfa, stems and leaves can be separated, and the leaves used as a high protein animal feed while the stems are used for fermentation or burning to produce energy (Delong et al., 1995, Lamb et al., 2007). A major advantage of developing alfalfa as a bioenergy crop over the other alternatives is that leaves could still be used as a high value commodity to supplement protein in livestock diet whereas only the low value stems will be converted to biofuels.

A significant component of the cost of production of many bioenergy crop is fossil fuel based synthetic nitrogen fertilizers, which can cause environmental problems (Patzek, 2004; Crews et al., 2004). Alfalfa can fix atmospheric nitrogen through its symbiosis with *Sinorhizobium meliloti*, and this can substantially decrease the cost of fertilizer inputs compared to non-leguminous crops. Further, being a perennial plant, alfalfa prevents soil erosion and provides significant wildlife habitat.

Diversity in Cell Wall Components

Understanding the structure of cell wall components and modifying them toward a desired level is vital for effective bioethanol production (Farrokhi et al., 2006). In alfalfa, the early efforts of understanding and manipulation of cell wall components were in the context of forage quality and digestibility (Buxton and Russell, 1987; Albrecht et al., 1988; Reddy et al 2005). Lignin is a phenolic compound and adds rigidity to plant structures. Forage legumes contain around 12% of lignin (Collins and Fritz, 2003). As cellulosic ethanol becomes important, the modification of cell wall constituents becomes more critical. Lignin inhibits forage digestibility in the same way as it reduces effective conversion of structural sugars into ethanol. The variability of lignin in alfalfa is a target of selection and forage nutritive value has been increased significantly by decreasing lignin (Shenk and Elliot, 1971; Kephart et al., 1989). Jung et al. (1997) evaluated the tetraploid alfalfa core collection for variation in forage quality and concluded that the core collection includes a significant amount of genetic variation.

Within a cell wall, cellulose is embedded in a matrix of lignin and hemicellulose, and pretreatment of biomass before fermentation is required in order for bacteria to have access to cellulose and hemicellulose and to convert them to fermentable sugars. This is one of the most expensive procedures during cellulosic bioethanol production (Dien et al., 2006). Down-regulation of enzymes in lignin biosynthesis has significantly lowered lignin content and increased digestibility in alfalfa (Baucher et al., 1999; Reddy et al, 2005; Gou et al., 2001a; Nakashima et al., 2000). The same approach has recently shown that reduction of lignin content could lead to effective saccarification at the pretreatment stage by increasing access to cellulose and hemicellulose (Chen and Dixon, 2007; Jackson et al., 2008).

Association Mapping

Most important agronomic traits are controlled by more than one gene. The genes underlying complex traits are termed Quantitative Trait Loci (QTL), and each has a different effect on the phenotype. The QTL that have relatively larger effects can be detected easily via genetic mapping. The goal of QTL mapping is to use Marker Assisted Selection (MAS) to facilitate the introgression of the QTL into desired breeding material and aid in its selection throughout the cultivar development process.

Different genetic mapping methodologies exist for detection of QTLs that control complex traits. Classical biparental mapping approaches depend on populations derived from two parents. While this process has been useful in identifying QTL for many traits in many crops (Wang et al., 1994; Bubeck et al., 1993), it has some limitations. First, since in diploids at most two alleles per locus can segregate in a cross from inbred lines, the amount of genetic variance within the population is limited (Malosetti et al. 2007), and much less than is present in most alfalfa breeding populations. Another restriction of this approach is the limited inference space of detected QTLs when considering the entire breeding populations (Jannink et al., 2001). Finally, in biparental mapping approaches, the maximum linkage disequilibrium (LD) is reached in the F_1 generation and with mapping usually occurring in the F_2 (or within relatively few numbers of generations after creating the F_1), LD has not decayed substantially, leading to large sections of chromosomes remaining in LD and consequently less precision in locating QTL. Consequently, in general, confidence intervals for QTL locations could be as high as 20 cM (Darvasi et al., 1993).

An alternative mapping approach is based on collections of genotypes from populations with a long history of recombination that decreases LD to (very) short chromosome intervals,

which is useful for creating strong and robust statistical marker-trait associations (Jannink et al., 2002). In association mapping, existing allele variation could be more efficiently represented and mapping could be directly conducted in the context of breeding material (Hirschhorn and Daly, 2005; Remington et al., 2001). Moreover, the precision with which a QTL can be located is much higher in association mapping compared to biparental mapping. However, if LD only extends over short distances, the biparental mapping approach is more powerful to detect the existence of a QTL, particularly if marker numbers are limited (Mackay and Powel, 2007).

Two major drawbacks exist in association mapping. First, false positive associations between markers and traits can be detected due to the presence of population structure (Aranza et al., 2005; Lander and Schork, 1994). However, population structure can be assessed with the marker information from genome wide genetic markers (such as SSRs), and association tests can then be conditioned on the population structure to reduce the false positive rate (Aranza et al., 2005; Pritchard et al., 2000). Second, the extent of LD plays a practical role for determining the number of markers needed for detection of association between genotype and phenotype. If the rate of LD decay is high, the association between loci and traits will be constrained to small physical distances. In such a situation, detection of association between genetic markers and the phenotype of interest requires more markers (Hagenblad and Nordborg, 2002). Hence, it is crucial to infer estimates of LD in the genome in order to effectively select a mapping strategy.

Candidate Gene Association

A genome wide association study requires a large number of markers to adequately cover the whole genome if LD decays quickly. Because the mating system of the species of interest and the population history affects LD, the number of markers needed is highly specific to the population of interest. Studies in humans have been conducted successfully using tens of

thousands of single nucleotide polymorphism (SNP) markers, but this level of marker saturation is not available for most agronomic crops at this time. When the limiting factor for association mapping is the absence of a large number of markers evenly dispersed throughout the genome, another strategy that can be used for detection of marker-trait associations is to assay variation in candidate genes (Neale, and Savolainen, 2004). Tracking genotypic variation via SNPs in possible candidate genes and associating this variation with phenotypic variation may be a better approach than using a low density of randomly chosen markers scattered throughout the genome. One of the disadvantages of this methodology is that it will not be able to detect possible QTLs residing on other parts of genome (Zhu et al., 2008). There are a number of successful candidate gene association studies reported recently (Thornsberry et al., 2001; Wilson et al., 2004; Weber et al., 2007; Weber et al., 2008). A suite of polymorphisms in the positional candidate gene *Dwarf8* was found to be associated with differences in flowering time in maize via the candidate gene association approach (Thornsberry et al., 2001). Wilson et al. (2004) used six maize candidate genes involved in kernel starch biosynthesis to asses association of genetic variation in these candidate genes with major kernel composition traits. Significant associations for kernel composition traits, starch pasting properties, and amylose levels were detected. Weber at al. (2007) used 48 markers from nine candidate regulatory genes to test if major regulatory genes of maize contribute to standing variation in Balsas teosinte and found ten associations involving five candidates (Weber et al., 2007). In a more comprehensive study, Weber et al. (2008) conducted a candidate gene association study in Balsas teosinte to test significant associations with 52 candidate genes against 31 phenotypic traits in five categories (flowering time, plant architecture, inflorescence architecture, kernel composition, and vegetative morphology) and found significant association for 10 traits in 15 candidate genes (Weber et al, 2008). Hence, the

candidate gene association approach is likely an effective tool for detecting novel marker trait associations when a genome wide association is not feasible.

REFERENCES

Albrecht, K. A., W. F. Wedin, and D. R. Buxton. 1987. Cell-wall composition and digestibility of alfalfa stems and leaves. *Crop Sci.* 27**,** 735–741.

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horto, K. Jakob, C. Lister, J. Molitor, C. Shindo, Tang C., et al. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1: e60.

Barnes, D.K., E.T. Bingham, R.P. Murphy, O.J. Hunt, D.F. Beard, W.H. Skrdla, and L.R. Teuber. 1977. Alfalfa germplasm in the United States: Genetic vulnerability, use, improvement, and maintenance. Tech. Bull. 1571. USDA-ARS, U.S. Gov. Print. Office, Washington, DC.

Barnes, R.F. and C.C. Sheaffer. 1995. Alfalfa p. 205-217 in Barnes et al. (ed) Forages (5^{th} edition) Iowa State Press.

Bauchan, G.R., T.A. Campbell, and M.A. Hossain. 2003. Comparative chromosome banding studies of nondormant alfalfa germplasm. Crop Science 43:2037–2042.

Baucher, M., M.A. BernardVailhe, B. Chabbert, J.M. Besle, C. Opsomer, M. VanMontagu, and J.Botterman. 1999. Down-regulation of cinnamyl alcohol dehydrogenase in transgenic alfalfa (*Medicago sativa* L.) and the effect on lignin composition and digestibility. *Plant Molecular Biology* 39: 437–447.

Brink, R.A., and D.C. Cooper. 1938. Partial self-incompatibility in *Medicago sativa* PNAS 24:497-499.

Brouwer, D.J. and T.C. Osborn. 1997. Identification of RFLP markers linked to the unifoliate leaf, cauliflower head mutation of alfalfa. *Journal of Heredity* 88, 150–152.

Brouwer, D.J., S.H. Duke, and T.C. Osborn. 2000. Mapping genetic factors associated with winter hardiness, fall growth, and freezing injury in tetraploid alfalfa. Crop Sci. 40:1387–1396.

Brown, R.C. 2003. Biorenewable resources. Iowa State Press, Ames, IA.

Brummer, E.C. 1999. Capturing heterosis in forage crop cultivar development. Crop Sci. 39:943–954.

Bubeck, D.M., M.M. Goodman, W.D. Beavis, D. Grant. 1993. Quantitative trait loci controlling resistance to gray leaf spot in maize. Crop Sci. 33: 838– 47.

Buxton, D. R. and J. R. Russell. 1988. Lignin constituents and cell-wall digestibility of grass and legume stems. Crop Sci. 28, 553–558.

Chen, F., and R.A. Dixon. 2007. Lignin modification improves fermentable sugar yields for biofuel production. Nature Biotechnology 25, 759-761.

Coburn, F. D. 1903. Alfalfa (*Medicago sativa*). Orange Judd Company. New York

Collins, M., and J.O. Fritz. 2003. Forage quality. p. 363–390. *In* R.F. Barnes et al. (ed.) Forages: An introduction to grassland agriculture. Iowa State Univ. Press, Ames, IA.

Crews, T.E. and M.B. Peoples. 2004. Legume versus fertilizer sources of nitrogen: Ecological tradeoffs and human needs. Agriculture, Ecosystems and Environment 102:279-297.

Darvasi, A., A. Weinreb, V. Minke, J. I. Weller and M. Soller. 1993. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134: 943–951.

Delong, M.M., D.R. Swanberg, E.A. Oelke, C. Hanson, M. Onischak, M.R. Schmid, and B.C. Wiant. 1995. Sustainable biomass energy production and rural economic development using alfalfa as a feedstock. p. 1582–1591. *In* D.L. Klass (ed.) Second Biomass Conf. of the Americas: Energy, Environment, Agriculture, and Industry, Portland, OR. 21–24 Aug.

Dien, B.S., H-J.G. Jung, K.P. Vogel et al. 2006. Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. Biomass Bioenergy 30:880–891.

Endre, G., P. Kaló, Z. Kevei, P. Kiss, S. Mihacea, B. Szakál, A. Kereszt, G.B. Kiss. 2002. Genetic mapping of the non-nodulation phenotype of the mutant MN-1008 in tetraploid alfalfa *Medicago sativa* Mol Gen Genet 266:1012–1019.

Farrokhi, N., R.A. Burton, L. Brownfield, M. Hrmova, S.M. Wilson, A. Bacic, G.B. Fincher. 2006. Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. Plant Biotechnology Journal 4:145–167.

Fehr W.R. 1987. Principles of cultivar development. Macmillan, New York.

Flajoulot, S., J. Ronfort, P. Baudouin, P. Barre, T. Huguet, C. Huyghe, B. Julier. 2005. Genetic diversity among alfalfa (*Medicago sativa*) cultivars coming from a breeding program, using SSR markers. Theor Appl Genet 111:1420–1429.

Flint-Garcia S. A., J. M. Thornsberry and E. S. Bucker, IV. 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54**:** 357–374.

Guo, D., F. Chen, K. Inoue et al. 2001. Down-regulation of Caffeic Acid 3-*O*-Methyltransferase and Caffeoyl CoA 3-*O*-methyltransferase in Transgenic Alfalfa (*Medicago sativa* L.):

Impacts on Lignin Structure and Implications for the Biosynthesis of G and S lignin. Plant Cell 13:73–88

Hagenblad, J. and M. Nordborg. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. *Genetics* 161:289–98.

Hanegraaf, M.C., E.E. Biewinga, C. van der Bijl. 1998. Assessing the ecological and economic sustainability of energy crops. Biomass and Bioenergy: 15(4/5):345–55.

Hedrick P. W. 1987. Gametic disequilibrium measures: proceed with caution *Genetics* 117, 331-341.

Hill, R.R., Jr., J.S. Shenk, and R.F Barnes. 1988. Breeding for yield and quality. p. 809–825. *In* A.A. Hanson et al. (ed.) Alfalfa and alfalfa improvement. ASA–CSSA–SSSA, Madison, WI.

Hirschhorn, J.N. and M.J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6: 95–108.

Hoisington, D., M. Khairallah, T. Reeves, J.M. Ribaut, B. Skovmand, S. Taba, and M.L. Warburton. 1999. Plant genetic resources: What can they contribute toward increased crop productivity? Proc. Natl. Acad. Sci. (USA) 96:5937–5943.

Holland, J.B., and E.T. Bingham. 1994. Genetic improvement for yield and fertility of alfalfa cultivars representing different eras of breeding. Crop Sci. 34:953–957.

Ivanov, A.I. 1988. Alfalfa. Amerind Publishing, New Delhi, India.

Jackson, L.A., G.L. Shadle, R. Zhou, J. Nakashima, F. Chen, and R.A. Dixon. 2008. Improving saccharification efficiency of alfalfa stems through modification of the terminal stages of monolignol biosynthesis. Bioenergy Research 1, 180-192.

Jannink, J. L., and B. Walsh. 2002. Association mapping in plant populations, pp. 59–68 in Quantitative Genetics, Genomics and Plant Breeding, edited by M. S. Kang. CAB International, New York.

Jung, H.G., C.C. Sheaffer, D.K. Barnes, and J.L. Halgerson. 1997. Forage quality variation in the U.S. alfalfa core collection. Crop Sci. 37:1361-1366.

Jung, H.J.G., and J.F.S. Lamb. 2004. Prediction of cell wall polysaccharide and lignin concentrations of alfalfa stems from detergent fiber analysis. Biomass & Bioenergy 27:365-373.

Kaló, P., G. Endre, L. Zimányi, G. Csanádi, and G.B. Kiss. 2000. Construction of an improved linkage map of diploid alfalfa (Medicago sativa). Theor Appl Genet. 100:641–657.

Kephart, K.D., D.R. Buxton, and R.R. Hill, Jr. 1989. Morphology of alfalfa divergently selected for herbage lignin concentration. Crop Sci. 29:778–782.

Kidwell, K.K., D.F. Austin, and T.C. Osborn. 1994. RFLP evaluation of nine *Medicago* accessions representing the original germplasm sources for North American alfalfa cultivars. Crop Sci. 34:230–236.

Lesins, K. & I. Lesins. 1979. Genus *Medicago* (Leguminasae): A taxogenetic Study. Kluwer, Dordrecht, Netherlands.

Lamb, J.F.S., C.C. Sheaffer, L.H. Rhodes, R.M. Sulc, D.J. Undersander, and E.C. Brummer. 2006. Five decades of alfalfa cultivar improvement: Impact on forage yield, persistence, and nutritive value. Crop Sci. 46:902–909.

Lander, E.S. and N.J. Schork. 1994. Genetic dissection of complex traits. Science 265: 2037–2048.

Liu, K.J., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128.

Mackay, I. and W. Powell. 2007. Methods for linkage disequilibrium mapping in crops. Trends in Plant Science Vol.12 No.2.

Malosetti, M., C. G. van der Linden, B. Vosman and F. A. van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato genetics 175: 879–889.

Maureira, I.J., F. Ortega, H. Campos, and T. C. Osborn. 2004. Population structure and combining ability of diverse Medicago sativa germplasms. Theor Appl Genet 109: 775–782.

McCoy, T.J. and Bingham E.T. 1988. Cytology and cytogenetics of alfalfa. p. 739-776, *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Mengoni, A., A. Gori, and M. Bazzicalupo. 2000. Use of RAPD and microsatellite (SSR) variation to assess genetic relationships among populations of tetraploid alfalfa, *Medicago sativa*. *PlantBreeding*. 119, 311–317.

Michaud, R., W.F. Lehman and M.D. Rumbaugh. 1988. World distribution and historical development in Alfalfa and alfalfa improvement. ASA, CSSA, and SSSA, Madison WI.

Nakashima, J., F. Chen, L. Jackson et al (2008) Multi-site genetic modification of monolignol biosynthesis in alfalfa (*Medicago sativa* L.): effects on lignin composition in specific cell types. New Phytol 179:738–750.

Neale, D.B. and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends Plant Sci. 9:325-330.

Patzek, T.W. 2004. Thermodynamics of the corn-ethanol biofuel cycle. Crit. Rev. Plant Sci. 23:519-567.

Pritchard, J.K., Stevens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pupilli, F., S, Businelli, F. Paolocci, C. Scotti, F. Damiani, and S. Arcioni. 1996. Extent of RFLP variability in tetraploid populations of alfalfa, *Medicago sativa*. Plant Breed 115:106-112.

Pupilli, F., P. Labombarda, C. Scotti, and S. Arcioni. 2000. RFLP analysis allows for the identification of alfalfa ecotypes. Plant Breed 119:271–276.

Quiros, C.F., and G.R. Bauchan. 1988. The genus *Medicago* and the origin of the *Medicago sativa* complex, p. 93-124, *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Ragauskas, A.J., C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, Jr., J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer, and T. Tschaplinski. 2006. The path forward for biofuels and biomaterials. Science 311:484-489.

Reddy, M.S.S., F. Chen, G. Shadle, L. Jackson, H. Aljoe, and R.A. Dixon. 2005. Targeted down-regulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (*Medicago sativa* L.). Proc Natl Acad Sci USA 102:16573–16578.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98: 11479–11484.

Riday, H. and E.C. Brummer. 2002a. Forage yield heterosis in alfalfa. Crop Sci 42:716-723

Riday, H. and E.C. Brummer. 2002b. Heterosis of agronomic traits in alfalfa. Crop Sci. 42:1081-1087

Robins, J.G., D. Luth, T.A. Campbell, G.R. Bauchan, C. He, D.R. Viands, J.L. Hansen, and E.C. Brummer. 2007a. Genetic mapping of biomass production in tetraploid alfalfa (Medicago sativa L.). Crop Sci. 47:1–10.

Robins, J.G., G.R. Bauchan, and E.C. Brummer. 2007b. Genetic mapping forage yield, plant height, and regrowth at multiple harvests in tetraploid alfalfa (*Medicago sativa* L.). Crop Sci. 47:11–18.

Rowlands, D.G. 1964. Self-incompatibility in sexually propagated cultivated plants. *Euphytica*, 13, 157.

Rumbaugh, M.D., J.L. Caddel and D.E. Rowe. 1988. Breeding and quantitative genetics in Alfalfa and alfalfa improvement, ASA, CSSA, and SSSA, Madison, WI.

Şakiroğlu, M. and E.C. Brummer. 2007. Little heterosis between alfalfa populations derived from the Midwestern and Southwestern United States. Crop Sci 47:2364-2371.

Segovia-Lerma, A., R.G. Cantrell, J.M. Conway, and I.M. Ray. 2003. AFLP-based assessment of genetic diversity among nine alfalfa germplasms using bulk DNA templates. Genome 46:51–58.

Shenk, J.S., and F.C. Elliott. 1971. Plant compositional changes resulting from two cycles of directional selection for nutritive value in alfalfa. Crop Sci. 11:521–524.

Sinskaya, E.N. 1961. Flora of Cultivated Plants of the U.S.S.R. XIII Perennial Leguminous. National Science Foundation, Washington D.C.

Sledge, M.K., J.H. Bouton, M. Dall'Agnoll, W.A. Parrott, and G. Kochert. 2002. Identification and confirmation of aluminum tolerance QTL in diploid *Medicago sativa* subsp. *coerulea*. Crop Sci. 42:1121–1128.

Sriwatanapongse S., and C.P. Wilsie. 1968. Intra and intervariety crosses of *Medicago sativa* L. and Medicago falcata L. Crop Sci. 8:465–466.

Tanksley, S.D. and S.R.McCouch 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277: 1063–1066.

Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler IV. 2001. Dwarf8 Polymorphisms associated with Variation in Flowering Time. Nat. Genet. 28, 286–289.

USDA, National Agricultural Statistical Service. 2004. State Level data for field crops: Hay http://www.nass.usda.gov:81/ipedb/main.htm

Vandemark, G.J., Ariss J.J., Bauchan G.A. et al. 2006. Estimating genetic relationships among historical sources of alfalfa germplasm and selected cultivars with sequence related amplified polymorphisms. Euphytica 152:9–16.

Volenec, J.J., S.M. Cunngingham, D.M. Haagenson, W.K. Berg, B.C. Joern, and D.W. Wiersma. 2002. Physiological genetics of alfalfa improvement: Past failures, future prospects. Field Crops Res. 75:97–110.

Wang, G. L., D. J. Mackill, J. M. Bonman, S. R. McCouch, M. C. Champoux, and R. J. Nelson. 1994. RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. Genetics 136: 1421-1434.

Westgate, J.M. 1910. Variegated alfalfa. USDA Bur. PI Ind. Bull. 169:1–63.

Waldron, L.R. 1920. First generation crosses between two alfalfa species. J. Am. Soc. Agron. 12:133–143.

Wilson, L.W., S. Whitt, A. Ibanez, T. Rocheford, M.M. Goodman, and E.S. Buckler, IV. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733.

CHAPTER 2

INFERRING POPULATION STRUCTURE AND GENETIC DIVERSITY OF A BROAD RANGE OF WILD DIPLOID ALFALFA (*MEDICAGO SATIVA* L.) ACCESSIONS USING SSR MARKERS[1]

[1] Sakiroglu, M., J.J. Doyle, and E. C. Brummer. To be submitted to *Theoretical and Applied Genetics*

ABSTRACT

Diversity analyses in alfalfa have mainly evaluated genetic relationships of cultivated germplasm, while the genetic basis of morphologically-based taxonomy, molecular diversity, and population structure of diploid germplasm in the *M. sativa-falcata* complex are largely unknown. A collection of 374 individual genotypes derived from 120 unimproved diploid accessions from the National Plant Germplasm System, including *M. sativa* subspecies *caerulea*, *falcata*, and *hemicycla*, were evaluated with 89 polymorphic SSR loci in order to estimate genetic diversity, infer the genetic bases of current morphology-based taxonomy, and determine population structure. High levels of variation were detected. A model-based clustering analysis of the genomic data identified two clearly discrete subpopulations, corresponding to the morphologically defined subspecies falcata and caerulea, with the hybrid subspecies hemicycla evident based on its genome composition representing a combination of the others. Two distinct subpopulations exist within each subsp. *caerulea* and subsp. *falcata*. Even though few individuals were evaluated per accession, over 60% of total genetic variation resided within accession. The results show that taxonomic relationships based on morphology are reflected in the genetic diversity data, and that clear distinctions among subspecies is evident at the diploid level.

INTRODUCTION

Alfalfa is grown on over 32 million hectares worldwide (Michaud et al. 1988). Cultivated alfalfa has been improved from a complex taxonomic group known as the *Medicago sativa-falcata* complex, which includes both diploid (2n=2x =16) and tetraploid (2n=4x=32) taxa. In addition to ploidy, flower color, pod shape, and pollen shape have traditionally been used to differentiate taxa. The diploid members of the complex include *M. sativa* subsp. *falcata*, with yellow flowers and sickle shaped pods, *M. sativa* subsp. *caerule*a, with purple flowers and pods with multiple coils, and their natural hybrid, *M. sativa* subsp. *hemicycla*, with variegated flower color and partially coiled pods. The tetraploid subspecies in the complex include *M. sativa* subsp. *sativa* (the direct analogue of diploid caerulea), *M. sativa* subsp. *falcata*, and the tetraploid hybrid *M. sativa* subsp. *varia*. (Quiros and Bauchan, 1988). Hybridization among taxa is possible even across ploidy levels by unreduced gametes (McCoy and Bingham, 1988). The genetic validity of current morphology-based classification of subspecies has not been confirmed.

The exploration of genetic diversity and population structure in alfalfa has generally focused on tetraploid breeding populations or progenitor germplasms. The nine progenitor germplasms that are accepted to be source of contemporary US alfalfa cultivars, *Medicago sativa* subspecies *falcata, Medicago sativa* subspecies *sativa* "Ladak", "Flemish", "Turkistan", "Indian", "African", "Chilean", "Peruvian", *and Medicago sativa* subspecies *varia*, (Barnes et al., 1977) have been studied. Kidwell et al. (1994) found that *Medicago sativa* subspecies *falcata* and Peruvian germplasm are distinct from the rest of germplasm. Segovia-Lerma et al. (2003) confirmed the distinctness of *Medicago sativa* subspecies *falcata* from the rest of historic germplasms. A similar study conducted in four of the nondormant germplasm, Indian, African, Chilean, and Peruvian via comparative chromosome banding revealed no separation between

nondormant germplasm except a week separation of Indian from the rest of germplasm (Bauchan et al., 2003). Comparison of some of the contemporary cultivars with the historical introductions indicated that Modern U.S. cultivars have deviated from historical germplasm over the time (Mauriera et al 2004; Vandermark et al., 2006). The effective distinction among narrow germplasm via molecular markers has been interest of breeders around the globe. Italian populations and ecotypes have been separated with RFLPs (Pupilli et al., 1996; Pupilli et al., 2000). The genetic diversity between Italian and Egyptian cultivars was evaluated with SSR and RAPD markers. The distinctness of Egyptian from the Italian cultivars was confirmed with both of the molecular marker types. Within population genetic variation of some Italian varieties and ecotypes evaluated with SSR markers was around 77% of total genetic variance (Mengoni et al. 2000). Brummer et al. (1991) used 19 cDNA RFLPs to infer population structure of alfalfa accessions via molecular markers and were able to differentiate *M. sativa* subspecies *falcata* from *M. sativa* subspecies *caerulea.* A comprehensive study of population structure across the breadth of natural populations and testing the consensus between current morphological taxonomy and molecular marker based phylogenies will help to determine if there is any rational of current morphology based taxonomy. It will also allow the evaluation of diploid gene pool of cultivated alfalfa. Although almost all the cultivated alfalfa is in tetraploid level, interploidy crosses allow the introgression of already available diploid germplasm. Evaluation and understanding of the population structure, allelic richness, and diversity parameters of diploid germplasm will help breeders to more effectively utilize genetic resources for cultivar development.

We selected a broad range of unimproved diploid *M. sativa* accessions from throughout the Northern Hemisphere. Our objectives in the present study were 1) to investigate the

population structure of a wide range of diploid members of the *M. sativa-falcata* species complex and concordance between current morphology based classification and differentiation based on SSR markers; 2) to infer the extent of genetic diversity that exists in diploid accessions.

MATERIALS AND METHODS

Sampling of Wild Diploid Alfalfa Populations

We obtained a large set of alfalfa accessions from the USDA National Plant Germplasm System. We used flow cytometry to identify diploid accessions. Ploidy was determined on a bulk of four genotypes per accession using previously described methods (Brummer et al., 1999) on a Cytomics FC 500 (Beckman-Coulter, Fullerton, CA**)** flow cytometer at the UGA Flow Cytometry Facility. If ploidy variation was observed in the bulked sample, each of the four genotypes was tested independently. As a result of the flow cytometry analysis, we selected 122 accessions and 384 genotypes to use in this experiment. Tetraploid accessions and individuals were not considered further. All ploidy information has been submitted to the USDA-NPGS.

The selected accessions represented the wide geographical distribution of *M. sativa* subsp. *caerulea,* hereafter referred to as caerulea, and *M. sativa* subsp. *falcata,* hereafter falcata, and their natural hybrid *M. sativa* subspecies *hemicycla,* hereafter hemicycla. The group of 122 accessions included 57 caerulea, 4 hemicycla, and 61 falcata according to their classification in the Germplasm Information Resource System (GRIN) when we began this experiment. Accessions were represented by between one and four individual genotypes, with 60 accessions having four individuals, 30 accessions having three, 16 accessions having two, and 15 accessions having one. One accession PI 641380 was represented by seven genotypes because it was included under its former designation (W6 4794) as well as its PI number, and we did not

discover the duplication until after samples had been analyzed. Seeds were germinated and grown in the greenhouse at Iowa State University and at the University of Georgia.

Because the primary character used to distinguish among the current taxa is flower color, we recorded flower color from each of the genotypes to check agreement of flower color with the classification listed in the Germplasm Resources Information Network (GRIN). Yellow flowered accessions were considered as falcata and purple flowered accessions as caerulea. Accessions with variegated flowers were listed as hemicycla. A secondary distinguishing characteristic is pod shape, with falcata having sickle shaped (falcate) pods and caerulea having coiled pods. We harvested pods from 50 caerulea, 50 falcata, and 21 hemicycla genotypes, recorded the coil number for 10 individual pods per genotype, and computed a mean coiling value. Pods were scored in ¼ coil intervals for the number of coils, ranging from completely straight (0 coils) to four coils.

DNA Extraction and SSR Genotyping

DNA was extracted from all 384 genotypes using young leaves from greenhouse grown plants, which were freeze-dried and ground to a powder. Genomic DNA was extracted following the CTAB method (Doyle and Doyle, 1989).

We selected 89 SSR markers from those used in previous studies in alfalfa (Diwan et al., 2000; Julier et al., 2003; Robins et al., 2007) that were easy to score across the diversity in this population. The M13 tailing method described by Schuelke (2000) was used to label PCR products. Each SSR marker was amplified by PCR independently using the protocol described by Julier et al. (2003) and Sledge et al. (2005). We pooled PCR products from 4-6 reactions for genotyping on an automated ABI3730 sequencer at the UGA DNA Sequencing Facility. Allele scoring was performed using GENEMARKER software (SoftGenetics, State College, PA) with

visual verification of all bands in all plants. Since we used diploid accessions, we scored genotypes in a biallelic genotypic format. The number of missing genotypes per marker was less than 5 percent. Based on marker profiles, ten of the 384 genotypes appeared to be tetraploid. We retested them with flow cytometry, confirmed their tetraploid status, and removed them from subsequent data analyses. Thus, all the results reported below are based on 374 individual genotypes derived from 120 accessions.

Data Analyses

In order to infer the population structure of the entire set of genotypes without regard to the preexisting subspecies classification or geographical information, we used the software program STRUCTURE (Pritchard et al., 2000). STRUCTURE provides a model-based approach to infer population structure by using our entire SSR marker dataset to identify *K* clusters to which it then assigns each individual genotype. Initially we used all 374 genotypes to deduce the true value of *K* (i.e., the number of clusters) by evaluating $K = 1$ to 10. In our model, admixture was allowed and the allele frequencies were assumed to be correlated. The length of burn-in Markov Chain Monte Carlo (MCMC) replications was set to 10,000 and data were collected over 100,000 MCMC replications in each run, based on previous literature suggesting that this level is sufficient (Evanno et al., 2005). The best estimation of the true number of *K* was obtained using the method developed by Evanno et al. (2005). We found that $K = 2$ results in the clear differentiation of two large groups, corresponding to falcata and caerulea, with a number of individual genotypes within those groups that show evidence of a hybrid genome, representing hemicycla. Because individual genotypes had varying proportions of the genome from each cluster, we developed an arbitrary classification as follows: 0-30% caerulea genome = falcata; 30-70% caerulea genome = hemicycla, and 70-100% caerulea genome = caerulea. A meaningful

population subdivision is also represented at K = 5, when each of the two large clusters corresponding to caerulea and falcata is subdivided into two groups, with the fifth cluster representing the hemicycla group.

In order to compare different ways of partitioning variance, we selected the 106 accessions for which two or more genotypes were evaluated and used them in an Analysis of Molecular Variance (AMOVA). We partitioned the genetic variance within and among accessions and within and among the five populations suggested by STRUCTURE. We conducted the AMOVA using the software program GenAlEx 6.1 (Peakall and Smouse, 2001).

In order to visualize the distribution of individual genotypes using the complete SSR dataset, a principal components analysis (PCA) was conducted using GenAlEx (Peakall and Smouse, 2001). A genetic distance matrix and neighbor joining cladogram were created using the software program TASSEL (Bradbury et al., 2007).

Several measures of diversity were computed, including the average number of alleles and genotypes per SSR locus, observed (H_0) and expected (H_e) heterozygosity, polymorphism information content, and the major allele frequency at each SSR locus. The computations were conducted over all genotypes and also by the subgroups identified above using the software program Power Marker v3.23 (Liu, 2002).

RESULTS

Morphological Analysis and Correction of Misclassifications

We initially defined the subspecies to which a particular accession belonged based on the classification in the GRIN system. However, accessions are occasionally misidentified in GRIN, so we clarified accession assignments based on the flower color data we recorded. Falcata are

defined as having yellow flowers, caerulea as having blue or purple flowers, and hemicycla as having variegated flowers (Lesins and Lesins, 1979; Quiros and Bauchan, 1988**)**. One of the goals of our experiment was to test the validity of flower color as distinguishing genetically discrete groups, so we reclassified accessions (and genotypes) according to their flower color prior to any further analysis.

Three accessions classified as falcata in GRIN (PI464726, PI464727, PI631814) have purple flowers and were reclassified as caerulea. The Afghan accession PI222198 was classified as caerulea in GRIN, but based on its yellow flower color, we reclassified it as falcata. One individual genotype from PI641603, which was initially classified as caerulea in GRIN, had variegated flower color and on that basis, we defined the entire accession as hemicycla. The taxonomic classification of these accessions was updated for our analysis below.

Pod shape is another morphological trait often used to differentiate alfalfa subspecies, with falcata accessions typically having sickle shaped pods with fewer than one coil, caerulea pods having multiple coils, and hemicycla falling in between (Quiros and Bauchan, 1988). The mean values of pod coils measured on our genotypes based on the taxonomic classification corrected for flower color showed that falcata has a mean of 0.2 coils per pod, hemicycla 1.1, and caerulea 2.0, as anticipated. These values are significantly different from one another based on least significant difference. However, nine accessions (PI307395, PI464728, PI577543, PI634119, PI634136, PI634174, PI641380, PI641601, and PI641603), had contradictory flower color and pod shape results, making unambiguous assignment to a particular subspecies impossible. For example, accession PI464728 had yellow flowers but a mean coil number per pod of 1.3; and the accession PI641601 had purple flowers and a mean coil number of 0.75.

Population Structure

Based on the second order statistics developed by Evanno et al. (2005) for STRUCTURE, the most likely true number of *K* was five. However, any value of *K* between two and five gave a biologically meaningful clustering. Assuming the true value of $K = 2$, the analysis essentially identified the two genomes corresponding to the falcata and caerulea subspecies (Figure 2.1A). Most of the individuals had greater than 70% of their genome derived from one subspecies; the remaining individuals, which had between 30 and 70% of their genomes from each group, were identified as hemicycla genotypes, showing very clearly the expected pattern of hybridity between the two subspecies (Figure 2.1A). The hierarchical population structure detected at $K = 5$ resulted in two subgroups nested within each of the falcata and caerulea subpopulations, breaking out the hemicycla group as a separate cluster (Figure 2.1B).

Based on our observation that the genomes of most individuals identified using $K = 2$ closely followed the morphologically based taxonomy, we used the genome percentages to classify all individuals. Those individuals with 0 to 30% caerulea genome were considered to be falcata; 31 to 69% hemicycla, and 70 to 100% caerulea. Based on this allocation, 44 individuals were placed into the hemicycla group. The three accessions (PI641615, PI641619, and PI634111) initially identified in GRIN as hemicycla are hybrids based on genome composition. All three individuals from PI641603 that we redefined as hemicycla based on flower color indicated a nearly even mixture of genomic backgrounds from each subspecies. The six caerulea accessions from Kazakhstan (PI634119, PI634136, PI634174, PI634176, PI641601, and 641606) and a Russian accession (PI315460) had purple flowers and were classified as caerulea in GRIN, but they had a hybrid genome pattern. This result suggests that the dominant purple flower color is not sufficient to identify accessions with hybrid genomes; these accessions should be

reclassified as hemicycla. Additionally, three of four individual genotypes belonging to the Georgian accession PI577543 showed a hybrid genome composition, suggesting that this accession should also be changed from caerulea to hemicycla. Some or all of the individual genotypes from accessions PI315460, PI577548, PI464727, PI464728 and PI631814 showed an interesting genome composition (Figure 2.1C). When *K*=2, the genotypes had genomes that were a mixture of falcata and caerulea. However, when *K*=5, the genome composition of the genotypes did not include the common "hemicycla" genome (colored pink in Fig. 1B), but rather consisted of hybrid patterns with differing amounts of genome composition from the four caerulea and falcata groups. Thus, while it appears these accessions are hemicycla, each genotype has a distinct genome composition. One individual genotype of the Russian accession PI577548 indicated a genome composition of approximately 50% from each of the two subspecies falcata and caerulea, but the other three genotypes of the accession had a genome of predominantly caerulea, suggesting the accession should remain classified as caerulea.

The reassignment of accessions based on genome composition resolved the flower color - pod shape discrepancies. All of the nine accessions that showed disagreement between the two morphological traits were reclassified as hemicycla. Based on the comparison of genome composition with flower color and pod shape, we found that yellow flowers and pods with one coil or more are a strong indication that an accession is hemicycla. All falcata accessions have yellow flowers and pods with less than one coil. The purple flowered accessions are harder to assign to either caerulea or hemicycla based on pod coiling. In general, caerulea accessions have pods with more than 1.5 coils. If hemicycla accessions have purple flowers, they tend to have pods with fewer than 1.5 coils.

Based on the morphological analysis and genome composition data, we assigned genotypes to subspecies and subgroups (Table 2.1). Of the 374 genotypes, 168 were caerulea, 162 falcata, and 44 hemicycla. Within the caerulea, group A had 99 genotypes and group B 69 genotypes. Falcata A had 100 genotypes and falcata B had 62 genotypes.

Principal Component Analyses and Neighbor-Joining Dendogram

We conducted a principal components analyses to further assess the population subdivisions identified using STRUCTURE. The first principal component explained 62% and the second principal component explained 11% of the SSR variation among the 374 genotypes. Plotting the first two principal components and color coding genotypes according to the five groups identified using STRUCTURE shows the clear separation of falcata and caerulea and the intermediate position of the hybrid hemicycla (Figure 2.2). Moreover, falcata accessions are clearly divided into two subgroups. The caerulea accessions form one large cluster, with the weak separation of the two subpopulations evident only by considering the identities of each individual (Figure 2.2).

We developed a neighbor-joining tree using TASSEL, which showed a pattern consistent with the two analyses above. Falcata and caerulea accessions are clearly separated and hemicycla genotypes show a clear hybrid pattern. The two falcata subgroups that are suggested by STRUCTURE are segregating in the tree as well. The segregation pattern of caerulea is also evident albeit less clearly (Figure 2.3). We also observed that the first falcata group (falcata A) if further subdivided into two clusters. The first cluster represents Russian falcata along with an accession from Sweden, whereas the second subgroup consists of all the European accessions that are allocated to falcata A. We also found that there is a cluster of genotypes from different subpopulations mainly collected from Eastern Russia or Eastern Kazakhstan (Figure 2.3).

Genotypes from the same accession are often but not exclusively in close proximity on the dendogram. Occasionally, genotypes of a given accession were placed in different subgroups of the same subspecies, as in the case of PI577558. The placement of different genotypes from the same accession into different groups was more common between the two caerulea subgroups, as expected. Differences in genome compositions were observed when different genotypes of the same accessions fell apart in the dendogram as was the case for PI315460. We also observed that the individuals that we denoted as hemicycla but that had a discrepancy in genome composition from other hemicycla accessions (Figure 2.1C) were placed with other subspecies. For example, one individual genotype from the accession PI577543 had around 65% of genome composed of caerulea and 35% from falcata and considered to be hemicycla, but it clustered with caerulea A, the subgroup with which it shared the largest portion of its genome (65%). A similar clustering pattern based on genome composition proportions was also observed in other genotypes.

AMOVA

We first conducted AMOVA among and within the five groups identified by our STRUCTURE analysis. In this analysis, 19% of the genetic variance was explained by group, with the remaining 81% residing within groups (Table 2.2). Next, only the 106 accessions that had 2-7 individuals were evaluated, partitioning genetic variation into within and among accession sources. Even with this restricted within accession sampling, 68% of the genetic variation was present within accessions with only 32% among accessions. Finally, we conducted a hierarchical AMOVA, partitioning variance among the five groups, among accessions within groups, and among genotypes within accessions. Genetic variance among groups was 19%, among accessions 16%, and within accessions 65% of the total (Table 2.2). Although most of the

genetic variance was among individuals within an accession, the differentiation of the accessions was also highly significant in all cases (p=0.001).

The pairwise Φ_{PT} values ranged from 0.059 (between hemicycla and caerulea B subgroups) to 0.294 (between caerulea A and falcata B) and each of the pairwise Φ_{PT} values was significantly different from zero according to tests based on 9999 random permutations ($P < 0.0001$) (Table 2.3).

Diversity Measurements

We evaluated diversity statistics for all 374 individual genotypes, for the three subspecies, and for the subgroups within subspecies. The overall number of alleles per SSR locus across 374 individuals ranged from 6 to 53 with a mean of 18.3 (Table 2.4). The mean number of genotypes per SSR locus for all 374 individuals was 52.8. Although caerulea had more genotypes than falcata (168 vs. 162), both the average number of alleles and average number of genotypes per SSR locus were higher in falcata. No obvious differences were evident among the subpopulations within either subspecies when taking the different number of individuals into account.

Overall observed heterozygosity ranged from 0.12 to 0.84 with a mean of 0.46. Gene diversity as a measurement of expected heterozygosity was higher in all cases compared to observed heterozygosity. Again as a measure of diversity, heterozygosity was higher in falcata than caerulea. The heterozygosity measure of hemicycla genotypes was 0.46 and showed an intermediate pattern of the other two subspecies (Table 2.4). The mean polymorphism level of loci for all genotypes was 0.71. The falcata group had a higher PIC value compare to caerulea, indicating more allelic diversity.

DISCUSSION

The classification of taxa in the *Medicago sativa* complex as species or subspecies has been controversial (Sinskaya, 1961; Lesins and Lesins, 1979; Ivanov, 1988; and Quiros and Bauchan, 1988). Sinskaya (1961) denoted caerulea, hemicycla, falcata, and *sativa* as species and considered them along with a number of other taxa as a "circle of species." Sinskaya (1961) first divided taxa based on ploidy, assuming that taxa within the same ploidy level were more closely related. Lesins and Lesins (1979) classified falcata and *sativa* as different species of the genus *Medicago*; however, hemicycla and caerulea were relegated to the subspecific level. They also noted that despite the obvious morphological distinctness of the two, there exist no hybridization obstacles between falcata and *sativa*. More recently, the previously defined species have been given subspecies status within the *M. sativa-falcata* complex (Quiros and Bauchan, 1988). Flower color and pod shape have been the primary characteristics used in taxonomic classification of members of the complex.

Based on different approaches to analyses genomic data, we found that the subspecies falcata and caerulea separated clearly with the hybrid nature of hemicycla observed based on its intermediate position between the other two subspecies. The separation between falcata and caerulea was observed in a narrow sampling of diploid accessions in a previous study using RFLPs (Brummer et al., 1991). Falcata and caerulea were also further subdivided into two groups. In order to interpret the separation within subspecies, we examined the habitat and geographical location from which each accession derived. Not all accessions have accurate location of collection information; however, we were able to use the available information to make a meaningful interpretation of our results. The two caerulea groups largely correspond to Northern and Southern regions of Eurasia (Figure 2.4). All Russian, Georgian, Armenian, and

Former Soviet Union accessions for which we do not have exact collection location information fall in the northern group (Figure 2.4). The accessions that were received from Canada but with an unknown original location of collection (PI571551 and PI571552) were also in the northern group. The southern group includes Turkish, Iranian, and Uzbekistani accessions and also represents the southern end of natural distribution of subspecies caerulea. Accessions from Kazakhstan were included in both groups depending on the location of collection. The accessions that were collected from northern Kazakhstan (Aqtobe) grouped with the northern cluster, whereas the accessions collected from southern Kazakhstan (Dzumbul) were allocated to the southern group. Three Kazakh accessions without exact collection location information (PI631922, PI631925, and PI577541) clustered with the northern caerulea group. Turkish accessions generally clustered with the southern group, but Turkish accession PI464726 grouped with the northern cluster (Figure 2.4).

The separation of falcata accessions into two groups more evident compared to the separation of caerulea based on the STRUCTURE analyses, but interestingly, this clear differentiation was not related to a pattern of geographic separation, as it was the case among caerulea accessions. Out of 33 accessions clustered in falcata A, GRIN provides the precise location of collection of 18 accessions. Fourteen of these accessions were collected from rocky mountain ranges or dry slopes. The two Mongolian accessions (PI641543 and PI641544) were collected from high elevations (732m and 762m respectively) around wheat fields. The Ukrainian accession PI634106 was collected from a small peninsula with elevation of around 100m; however, from a habitat described as a north slope, moderately steep, and on cliff. The Russian accession of PI631666 was collected from a moist stream terrace. We have precise collection location information for 13 of 19 accessions in falcata B. Of those 13 accessions, six

were collected from locations around rivers or wetlands, two accessions were gathered from coastal areas, two were from areas surrounded by dense forest, and three were collected from low elevation areas currently inhabited by humans.

Five accessions showed a hybrid genome pattern between the two falcata groups (PI577558, PI538987, PI631568, PI577555, and PI631707). The Italian accession (PI631568) was collected from a valley between a lowland plain and the Alps Mountains and showed a hybrid pattern based on genome composition. The two Russian accessions PI577558 and PI538987 were collected from the border of a river basin and dry plains (Figure 2.4). The Ukrainian accession PI577555 that showed hybrid pattern was collected from 100m elevation plain. The Russian accession PI631812 had one individual genotype assigned to falcata B whereas the other two genotypes assigned to falcata A.

Based on this information, we can infer that falcata A was collected from dry slopes of mountain ranges and/or high elevations. Therefore, we denoted them as highland ecotypes. In contrast, falcata B was collected from lowlands and locations close to river basins or the sea coast, so we named this group lowland ecotypes. In her effort to allocate former USSR falcata accessions into ecotypes, Sinskaya (1961) used a combination of geography and ecology to differentiate among them. In addition to designating the regions from which accessions derived, she mentioned five ecotypes as additional descriptors: floodland, steppe, submontane, mountain, and forest-steppe. She also connected different ecotypes based on root system development. We did not find evidence for all the differentiation that Sinskaya (1961) identified, but broad categories were recognized.

The USDA uses the same nomenclature of Quiros and Bauchan (1988) to denote taxa. We found that flower color is a powerful tool to assign individual genotypes to either caerulea or

falcata subspecies, but in some cases, it is not sufficient. The genome of particular individuals may contain a degree of admixture from another subspecies, so genome composition data can facilitate further assignments of individuals or accessions into subspecies hemicycla. Since the degree of admixture is continuous, we arbitrarily defined a cut-off point to determine if an individual genotype was hybrid or not. If less than 70% of the genome is estimated to belong to one of the subspecies (falcata or caerulea*)* then it has been defined as hemicycla, regardless of the flower color. At this point we can conclude that in order to accurately define a genotype, genomic data needed for estimation of the genomic composition. The individual genotypes we denoted as hemicycla based on genomic composition and which were originally classified as either falcata or caerulea based on flower color were from two regions of sympatry between caerulea and falcata. The first location is in Kars province of extreme northeastern Turkey. The other location is Aqtobe Kazakhstan, where almost the entire USDA hemicycla collection was gathered. Most of the accessions that have been denoted as hemicycla in GRIN are in fact tetraploid, and hence most likely belong to *M. sativa* subsp. *varia*. Extensive gene flow between the subspecies could occur in these regions because no known hybridity barriers between falcata and caerulea have been identified and artificial hybridization between them is routine.

Based on AMOVA, most of genetic variation in diploid alfalfa is within populations even when only a few individuals were used. With just two to four individual per accession, 68 percent of the genetic variance is within population. These results are broadly congruent with tetraploid alfalfa studies (e.g., Mengioni etc.) and more generally, with other outcrossing plants.

Unfortunately, we do not have the ability to directly compare our genetic diversity results with previous studies mainly because this is the first extensive study in diploids and the most comprehensive evaluation of genetic diversity in unimproved germplasm. Previous genetic

diversity assessment studies in alfalfa were conducted mainly in tetraploids, specifically to quantify the genetic diversity in relatively narrow breeding material. When we compare the findings of the current study with the similar analyses in tetraploid accessions, we found that the average number of alleles per SSR marker is much higher. This is expected since most earlier studies have been conducted within relatively narrow germplasm and on few individuals. For example, Fajoulot at al. (2005) found that the number of alleles per SSR locus ranged from 3 to 24 with a mean of 14.9. Another study conducted to determine the genetic diversity of Iranian cultivated local populations using SSRs revealed that the number of alleles per locus ranged from 6 to 11 with a mean of 9.2 (Bahar et al, 2006).

Based on this comprehensive genomewide SSR study of diploid members of the *M. sativa-falcata* complex, we can conclude that that the taxa considered here (falcata, caerulea, and hemicycla) are distinct to an extent that they could be denoted as subspecies. The hybrid nature of subspecies hemicycla was also very clear. We have also found that an immense amount of genetic variation exists in diploid accessions of cultivated alfalfa that could be used to broaden the genetic variation of cultivated alfalfa. The different ecotypes imply distinction of germplasm and this could help breeders to more effectively use diploid germplasm. Of particular importance, however, is the clarity with which caerulea and falcata are distinguished. Because these subspecies are apparently "pure," genome analysis of alfalfa should start with this material, determine if differences are present between the two diploid subspecies, and then compare those results with more complicated and intermixed tetraploid genotypes.

REFERENCES

Bahar, M., S. Ghobadi, V. Erfani Moghaddam, A. Yamchi, M. Talebi Bedaf, M. M. Kaboli, and A. A. Mokhtarzadeh. 2006. Evaluating genetic diversity of Iranian alfalfa local populations using expressed sequence tags (ESTs) microsatellites. J. Sci. & Technol. Agric. & Natur. Resour., Vol. 10, No. 2, Summer 2006, Isf. Univ. Technol., Isf., Iran.

Barnes, D.K., E.T. Bingham, R.P. Murphy, O.J. Hunt, D.F. Beard, W.H. Skrdla, and L.R. Teuber. 1977. Alfalfa germplasm in the United States: Genetic vulnerability, use, improvement, and maintenance. Tech. Bull. 1571. USDA-ARS, U.S. Gov. Print. Office, Washington, DC.

Bauchan, G.R., T.A. Campbell, and M.A. Hossain. 2003. Comparative chromosome banding studies of nondormant alfalfa germplasm. Crop Science 43:2037–2042.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ram-Doss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Brummer, E. C. 2004. Genomic research in alfalfa (Medicago sativa L.) Legume crop genomics. R. F. Wilson, H. T. Stalker and E. C. Brummer. Champaign, Illinois, AOCS Press: 110-141.

Brummer, E.C., G. Kochert, and J.H. Bouton. 1991. RFLP variation in diploid and tetraploid alfalfa Theor. Appl. Gen. 83:89-96.

Diwan, N., J.H. Bouton, G. Kochert, and P.B. Cregan. 2000. Mapping of simple sequence repeat (SSR) DNA markers in diploid and tetraploid alfalfa. Theor. Appl. Genet. 101:165–172.

Doyle, J. & J. Doyle. 1989. Isolation of plant DNA from fresh tissue. Focus life technologies 12: 1.

Evanno, G., S. Regnaut, J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611–2620.

Flajoulot, S., J. Ronfort, P. Baudouin, P. Barre, T. Huguet, C. Huyghe, B. Julier. 2005. Genetic diversity among alfalfa (*Medicago sativa*) cultivars coming from a breeding program, using SSR markers. Theor Appl Genet 111:1420–1429.

Ivanov, A.I. 1988. Alfalfa. Amerind Publishing, New Delhi, India.

Julier, B., S. Flajoulot, P. Barre, G. Cardinet, S. Santoni, T. Huguet, and C. Huyghe. 2003. Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago sativa*) using microsatellite and AFLP markers. BMC Plant Biol. 3:9.

Lesins, K. & I. Lesins. 1979. Genus *Medicago* (Leguminasae): A taxogenetic study. Kluwer, Dordrecht, Netherlands.

Liu, J. 2002. POWERMARKER – A powerful software for marker data analysis. Raleigh, NC: North Carolina State University, Bioinformatics Research Center. (http://www.powermarker.net).

Maureira, I.J., F. Ortega, H. Campos, and T. C. Osborn. 2004. Population structure and combining ability of diverse Medicago sativa germplasms. Theor Appl Genet 109: 775–782.

McCoy, T.J. and Bingham E.T. 1988. Cytology and Cytogenetics of Alfalfa. p. 739-776, *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA--SSSA, Madison, WI.

Mengoni, A., A. Gori, and M. Bazzicalupo. 2000. Use of RAPD and microsatellite (SSR) variation to assess genetic relationships among populations of tetraploid alfalfa, *Medicago sativa*. *PlantBreeding*. 119, 311–317.

Michaud, R., W. F. Lehman, M.D. Rumbaugh. 1988. World Distribution and Historical Development. P.25-91. In A. A. Hanson, D.K. Barnes, and R.R. Hill (ed.). Alfalfa and Alfalfa Improvement. ASA-CSSA-SSSA, Madison, WI.

Peakall, R. and P.E. Smouse. 2001. *GenAlEx V6: genetic analysis in Excel*. Population genetic software for teaching and research. Australian National University, Canberra . http://www.anu.edu.au/BoZo/GenAlE.

Pritchard, J.K., Stevens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pupilli, F., S, Businelli, F. Paolocci, C. Scotti, F. Damiani, and S. Arcioni. 1996. Extent of RFLP variability in tetraploid populations of alfalfa, *Medicago sativa*. Plant Breed 115:106-112.

Pupilli, F., P. Labombarda, C. Scotti, and S. Arcioni. 2000. RFLP analysis allows for the identification of alfalfa ecotypes. Plant Breed 119:271–276.

Quiros, C.F., and G.R. Bauchan. 1988. The genus *Medicago* and the origin of the *Medicago sativa* complex, p. 93-124, *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Robins, J.G., D. Luth, T.A. Campbell, G.R. Bauchan, C. He, D.R. Viands, J.L. Hansen, and E.C. Brummer. 2007. Genetic mapping of biomass production in tetraploid alfalfa (Medicago sativa L.). Crop Sci. 47:1–10.

Schuelke, M. 2000. An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234.

Segovia-Lerma. A., R.G. Cantrell, J.M. Conway, and I.M. Ray. 2003. AFLP-based assessment of genetic diversity among nine alfalfa germplasms using bulk DNA templates. Genome 46:51–58.

Sinskaya, E.N. 1961. Flora of cultivated plants of the U.S.S.R. XIII Perennial Leguminous. National Science Foundation, Washington D.C.

Sledge, M.K., I.M. Ray, and G. Jiang. 2005. An expressed sequence tag SSR map of tetraploid alfalfa (*Medicago sativa* L.). Theor. Appl. Genet. 111:980–992.

Vandemark, G.J., Ariss J.J., Bauchan G.A. et al. 2006. Estimating genetic relationships among historical sources of alfalfa germplasm and selected cultivars with sequence related amplified polymorphisms. Euphytica 152:9–16.

Table 2.1. List of all the accessions used in this study along with the number of individual genotypes used, country of origin, number of chromosomes, flower color, and classification of each accession based on this study

PI Number	Number of Genotypes Used	Country of Origin	Chromosome Number (2n)	Flower Color	Classification	
PI 179370	4	Turkey	16	Purple	caerulea	A
PI 210367	4	Iran	16	Purple	caerulea	A
PI 212798	4	Iran	16	Purple	caerulea	A
PI 222198	1	Afghanistan	16	Yellow	falcata[a]	B
PI 243225	4	Iran	16	Purple	caerulea	A
PI 251690	1	Former Soviet Union	16	Yellow	falcata	B
PI 251830	3	Austria	16	Yellow	falcata	A
PI 258752	2	Russia	16	Yellow	falcata	B
PI 283640	4	Former Soviet Union	16	Purple	caerulea	A
PI 299045	4	Russian Federation	16	Purple	caerulea	B
PI 299046	4	Russian Federation	16	Purple	caerulea	B
PI 307395	4	Former Soviet Union	16	Purple	caerulea	B
PI 314267	4	Uzbekistan	16	Purple	caerulea	A
PI 314275	4	Uzbekistan	16	Purple	caerulea	A
PI 315460	2	Russian Federation	16	Purple/Variegated	hemicycla[a]	
PI 315462	1	Russian Federation	16	Purple	caerulea	B
PI 315466	4	Russian Federation	16	Purple	caerulea	B
PI 315480	4	Russia	16	Yellow	falcata	B
PI 325387	2	Russia	16	Yellow	falcata	B
PI 325396	3	Russia	16	Yellow	falcata	B
PI 325399	3	Russia	16	Yellow	falcata	B
PI 440500	4	Kazakhstan	16	Purple	caerulea	A
PI 440501	4	Kazakhstan	16	Purple	caerulea	A
PI 440502	4	Kazakhstan	16	Purple	caerulea	A
PI 440505	4	Kazakhstan	16	Purple	caerulea	A
PI 440507	4	Kazakhstan	16	Purple	caerulea	A
PI 440514	2	Kazakhstan	16	Purple	caerulea	A
PI 464712	4	Turkey	16	Purple	caerulea	A
PI 464713	4	Turkey	16	Purple	caerulea	A
PI 464714	3	Turkey	16	Purple	caerulea	A
PI 464715	4	Turkey	16	Purple	caerulea	A
PI 464717	3	Turkey	16	Purple	caerulea	A
PI 464718	4	Turkey	16	Purple	caerulea	A
PI 464719	4	Turkey	16	Purple	caerulea	A
PI 464720	1	Turkey	16	Purple	caerulea	A
PI 464721	4	Turkey	16	Purple	caerulea	A
PI 464722	4	Turkey	16	Purple	caerulea	A
PI 464723	2	Turkey	16	Purple	caerulea	A
PI 464724	3	Turkey	16	Purple	caerulea	A
PI 464726	3	Turkey	16	Purple	caerulea[b]	A
PI 464727	2	Turkey	16	Purple	hemicycla[b]	
PI 464728	3	Turkey	16	Yellow	hemicycla[b]	
PI 464729	2	Turkey	16	Yellow	falcata	B
PI 486205	3	Russia	16	Yellow	falcata	B
PI 486206	3	Russia	16	Yellow	falcata	B
PI 486207	4	Russia	16	Yellow	falcata	B
PI 494662	4	Romania	16	Yellow	falcata	A
PI 502425	1	Russia	16	Purple	caerulea	B

PI 502437	3	Russian Federation	16	Purple	caerulea	B
PI 502438	1	Russia	16	Yellow	falcata	B
PI 502447	3	Russia	16	Yellow	falcata	A
PI 502448	4	Russia	16	Yellow	falcata	A
PI 502449	4	Former Soviet Union	16	Yellow	falcata	A
PI 505871	3	Former Soviet Union	16	Purple	caerulea	B
PI 538987	2	Russia	16	Yellow	falcata	A
PI 577541	3	Kazakhstan	16	Purple	caerulea	B
PI 577543	4	Georgia	16	Purple	hemicycla[a]	
PI 577545	1	Russian Federation	16	Purple	caerulea	B
PI 577546	4	Georgia	16	Purple	caerulea	B
PI 577547	4	Georgia	16	Purple	caerulea	B
PI 577548	4	Russian Federation	16	Purple	caerulea	B
PI 577549	2	Georgia	16	Purple	caerulea	A
PI 577551	4	Canada	16	Purple	caerulea	B
PI 577552	2	Canada	16	Purple	caerulea	B
PI 577555	4	Ukraine	16	Yellow	falcata	B
PI 577556	4	Bulgaria	16	Yellow	falcata	B
PI 577558	3	Russia	16	Yellow	falcata	A
PI 577564	2	Russia	16	Yellow	falcata	A
PI 631546	3	Russia	16	Yellow	falcata	B
PI 631549	2	Russia	16	Yellow	falcata	B
PI 631556	1	Russia	16	Yellow	falcata	B
PI 631561	4	Switzerland	16	Yellow	falcata	A
PI 631566	2	Bulgaria	16	Yellow	falcata	A
PI 631568	4	Italy	16	Yellow	falcata	A
PI 631571	2	Bulgaria	16	Yellow	falcata	A
PI 631577	2	Italy	16	Yellow	falcata	A
PI 631650	4	Bulgaria	16	Yellow	falcata	A
PI 631652	4	Russia	16	Yellow	falcata	B
PI 631654	3	Russia	16	Yellow	falcata	B
PI 631656	3	Russia	16	Yellow	falcata	B
PI 631658	3	Russia	16	Yellow	falcata	B
PI 631660	4	Russia	16	Yellow	falcata	B
PI 631661	3	Russia	16	Yellow	falcata	B
PI 631666	4	Russia	16	Yellow	falcata	B
PI 631667	4	Russia	16	Yellow	falcata	B
PI 631668	4	Russia	16	Yellow	falcata	B
PI 631689	3	Bulgaria	16	Yellow	falcata	A
PI 631691	2	Bulgaria	16	Yellow	falcata	A
PI 631707	1	China	16	Yellow	falcata	B
PI 631807	3	Russia	16	Yellow	falcata	A
PI 631808	1	Russia	16	Yellow	falcata	A
PI 631809	2	Russia	16	Yellow	falcata	A
PI 631812	3	Russia	16	Yellow	falcata	B
PI 631813	4	Russia	16	Yellow	falcata	A
PI 631814	3	Russia	16	Purple	hemicycla[b]	
PI 631816	4	Russia	16	Yellow	falcata	B
PI 631817	2	Russia	16	Yellow	falcata	B
PI 631818	3	Russia	16	Yellow	falcata	A
PI 631829	3	Russia	16	Yellow	falcata	B
PI 631842	1	Sweden	16	Yellow	falcata	A
PI 631921	4	Russian Federation	16	Purple	caerulea	B
PI 631922	4	Kazakhstan	16	Purple	caerulea	B
PI 631924	4	Armenia	16	Purple	caerulea	A
PI 631925	4	Kazakhstan	16	Purple	caerulea	B
PI 631926	4	Russian Federation	16	Purple	caerulea	B
PI 634034	1	Russia	16	Yellow	falcata	B
PI 634106	2	Ukraine	16	Yellow	falcata	B
PI 634119	4	Kazakhstan	16	Purple	hemicycla[a]	
PI 634136	4	Kazakhstan	16	Purple	hemicycla[a]	
PI 634174	4	Kazakhstan	16	Purple	hemicycla[a]	

PI 634176	3	Kazakhstan	16	Purple	hemicycla[a]	
PI 641380	7	Russian Federation	16	Purple/Variegated	caerulea	B
PI 641543	4	Mongolia	16	Yellow	falcata	B
PI 641544	4	Mongolia	16	Yellow	falcata	B
PI 641601	4	Kazakhstan	16	Purple/Variegated	hemicycla[a]	
PI 641603	3	Kazakhstan	16	Purple/Variegated	hemicycla[a]	
PI 641606	3	Kazakhstan	16	Purple	hemicycla	
PI 641615	4	Kazakhstan	16	Purple	hemicycla	
PI 641619	1	Kazakhstan	16	Purple	hemicycla	
SD201	1		16	Yellow	falcata	B

[a] Initially defined as caerulea in USDA-GRIN

[b] Initially defined as falcata in USDA-GRIN

Table 2.2. AMOVA tables of three different ways of analyzing the molecular variance of 362 genotypes of 106 accessions belonging to five different groups.

Source	df	SS	MS	Variance Component	% Total	φ	P
Among Groups	4	5468	1367	18	19%	0.191	0.0001
Within Groups	355	27435	78	77	81%		
Among Accessions	105	17093	163	29	32%	0.321	0.0001
Within Accessions	254	15810	65	62	68%		
Among Groups	4	5468	1367	18	19%	0.185	0.0001
Among Accessions	101	11625	115	15	16%	0.199	0.0001
Within Accessions	254	15810	62	62	65%	0.347	0.0001
Total	359	32904		95	100%		

Table 2.3. Pairwise Φ_{PT} values of the five groups detected based on STRUCTURE analysis.

Group	Caerulea A	Caerulea B	Hemicycla	Falcata A
Caerulea B	0.063			
Hemicycla	0.112	0.059		
Falcata A	0.283	0.239	0.144	
Falcata B	0.294	0.241	0.148	0.083

All ΦPT values are significant at P=0.0001

Table 2.4. Means along with ranges (in the parenthesis) of diversity statistics based on 89 SSR loci of 374 individual genotypes of subspecies caerulea, falcata, hemicycla, and the subgroups.

Parameter	Overall	*M. sativa* subsp. *caerulea*			*M. sativa* subsp. *falcata*			*M. sativa* subsp. *hemicycla*
		overall	group A	group B	overall	group A	group B	
Number of individuals	374	168	99	69	162	100	62	44
Allele No.	19.3 (6 – 53)	13.67 (3-41)	10.85 (3-31)	10.54 (3-32)	15.20 (4-46)	12.90 (3-39)	11.06 (3-28)	9.7 (2-21)
Genotype No.	52.8 (10-149)	28.92 (4-82)	20.76 (3-58)	18.24 (4-52)	34.63 (7-103)	25.60 (6-77)	19.46 (3-50)	15.4 (3-31)
Major Allele Frequency	0.39 (0.13-0.94)	0.50 (0.14-0.93)	0.52 (0.13-0.98)	0.50 (0.13-0.91)	0.43 (0.10-0.90)	0.45 (0.09-0.94)	0.45 (0.15-0.96)	0.44 (0.14-0.90)
Gene Diversity	0.74 (0.29-0.94)	0.65 (0.11-0.92)	0.62 (0.02-0.91)	0.64 (0.18-0.92)	0.70 (0.20-0.94)	0.68 (0.12-0.95)	0.68 (0.08-0.93)	0.69 (0.19-0.92)
Heterozygosity	0.46 (0.12-0.81)	0.44 (0.07-0.80)	0.43 (0.02-0.83)	0.46 (0.09-0.83)	0.48 (0.11-0.81)	0.50 (0.06-0.90)	0.45 (0.08-0.89)	0.46 (0.07-0.84)
PIC	0.71 (0.28-0.94)	0.62 (0.11-0.92)	0.59 (0.03-0.91)	0.61 (0.17-0.91)	0.67 (0.19-0.94)	0.65 (0.12-0.94)	0.65 (0.08-0.93)	0.66 (0.19-0.91)

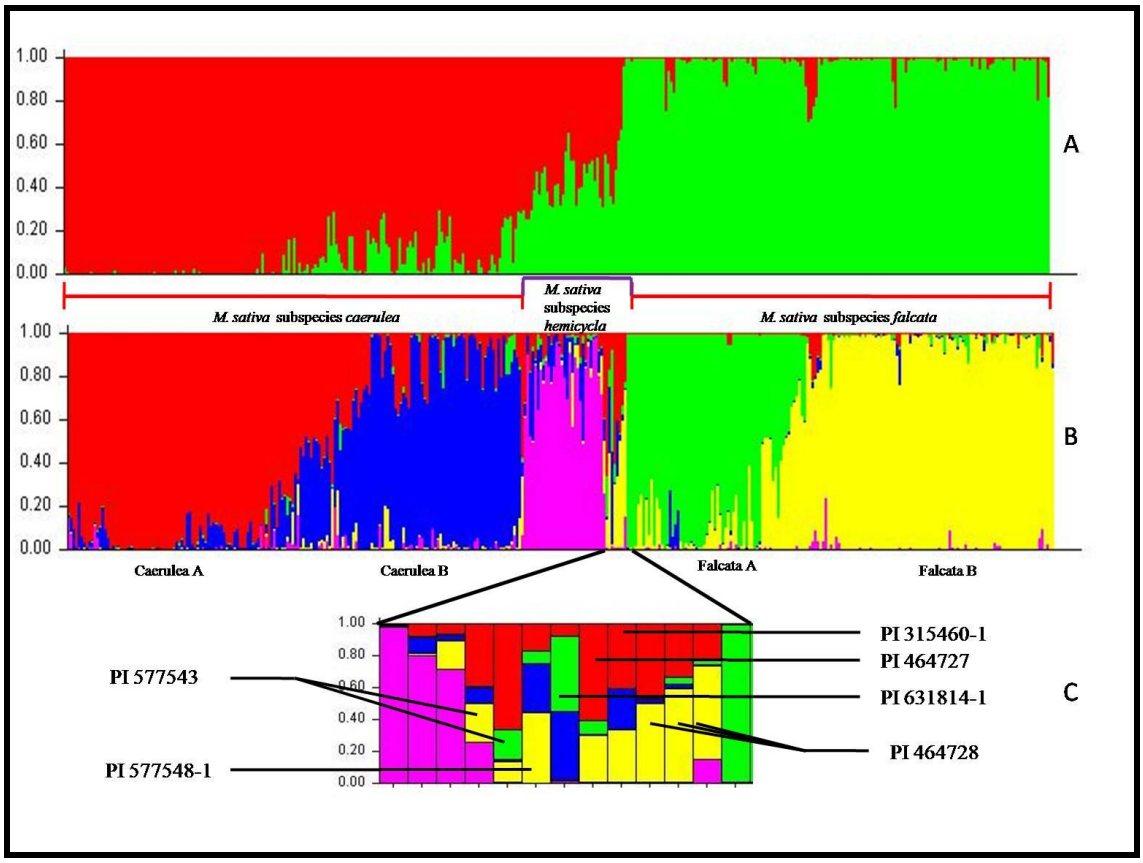

Figure 2.1. Identification of the possible populations in the data sets. A) 374 individual genotypes when the true value of K is estimated to be 2. The subpopulation that is represented predominantly by red color is denoted as caerulea, whereas green represents the other subpopulation that contains falcata accessions B) Separation of subpopulations when K is considered to be five. C. The hemicycla genotypes that deviate from the rest of the subpopulation in terms of genome composition and their corresponding PI number.

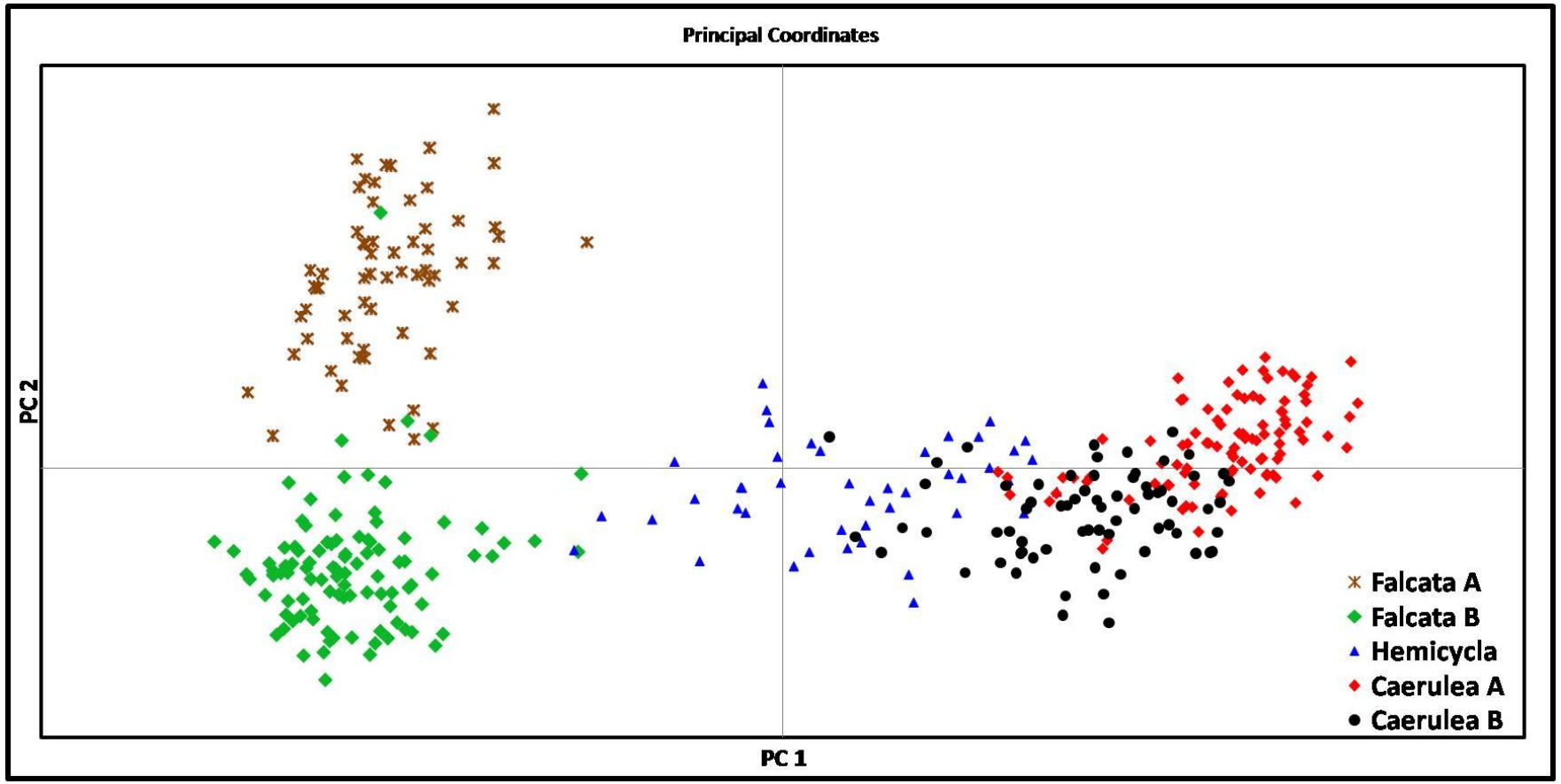

Figure 2.2. Differentiation of the five populations based on the first two principal components.

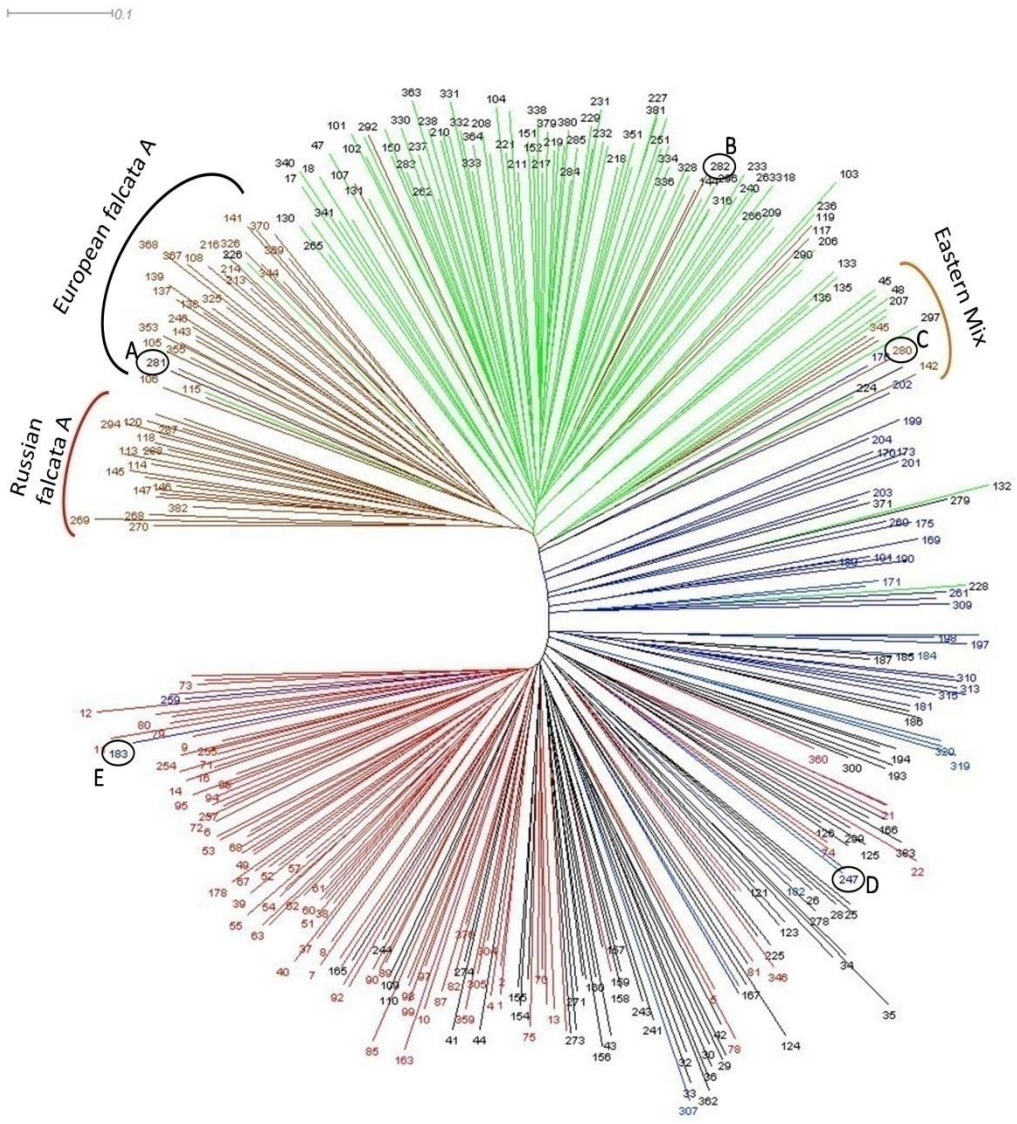

Figure 2.3: NJ tree of 374 individual genotypes of wild unimproved diploid accession of *M. Sativa* L. Green Color represents falcata B (UF) genotypes and brown represents falcata A (LF). Blue color represents hemicycla genotypes. Red color represents caerulea A (SC) and black represent caerulea B (NC) genotypes. The three genotypes labeled A, B, and C represent three different genotypes of accession of PI577558 that are placed in different clusters. D represents the two genotypes of the accession PI315460 that is defined as hemicycla but grouped with caerulea. E represents an individual genotype from PI577543.

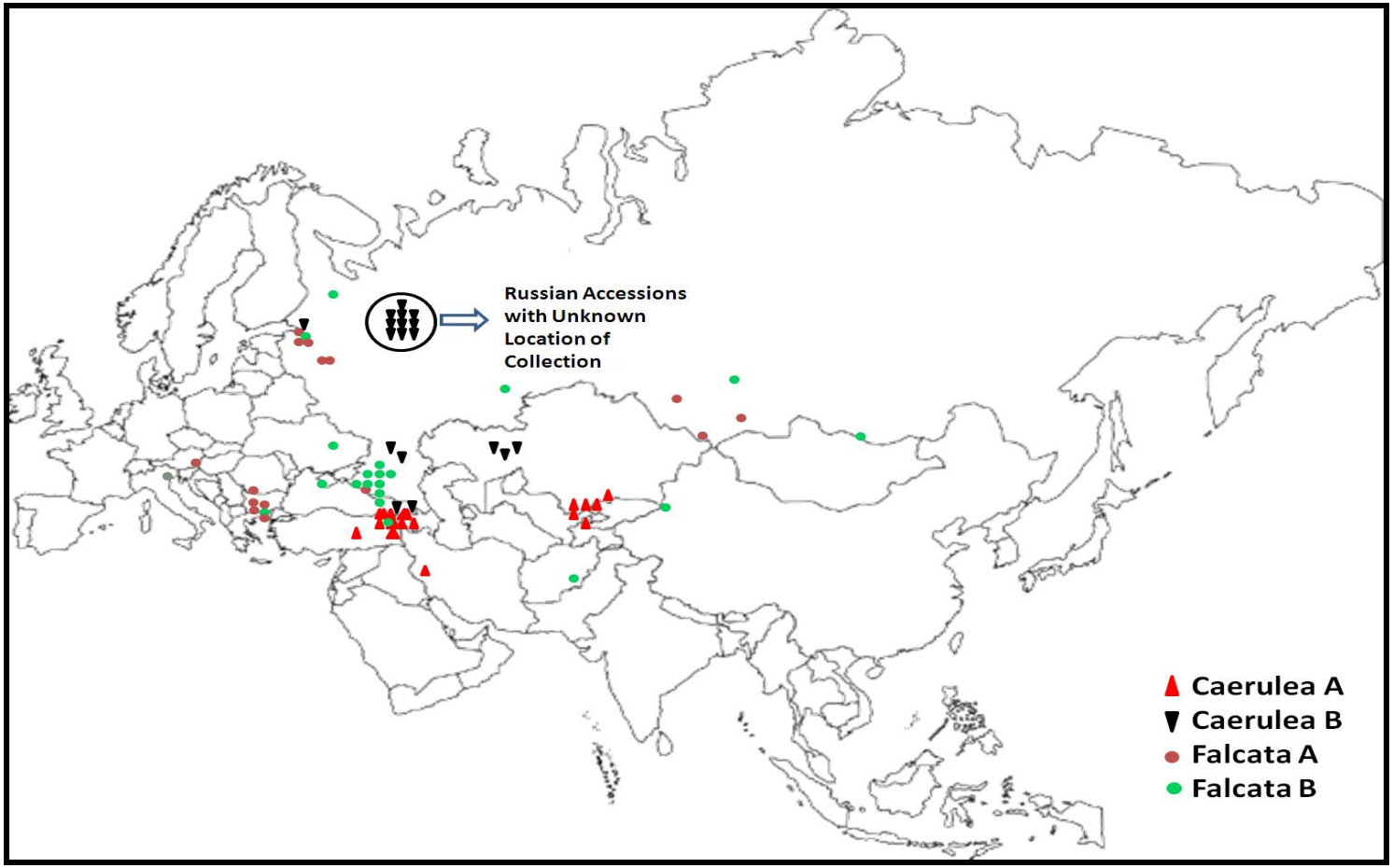

Figure 2.4. The map of collection of locations of the two caerulea subgroups and two falcata subgroups.

CHAPTER 3

VARIATION IN BIOMASS YIELD, CELL WALL COMPONENTS, AND AGRONOMIC TRAITS IN A BROAD RANGE OF DIPLOID ALFALFA (*M. SATIVA* L.*)* ACCESSIONS[1]

[1] Sakiroglu, M., K.J. Moore, and E.C. Brummer. To be submitted to *Crop Science*

ABSTRACT

Alfalfa is an important forage and a potential biofuel crop, but breeding is needed to improve its yield and composition. Agronomically useful genetic variation may be present in the diploid germplasm pool, which has previously been little used in alfalfa breeding. We gathered 374 individual genotypes from 120 accessions of wild diploid alfalfa collected from throughout the Northern hemisphere and evaluated their performance for cell wall constituents, total biomass yield, and other agronomic traits in field trials at two Georgia environments in 2007 and 2008. A large amount of phenotypic variation exists among diploid accessions for agronomic traits and for cell wall constituents. No particular population structure based on these agronomic characteristics was observed unlike clear differentiation of subspecies based on genetic data. The results show that diploids are a potentially useful pool of alleles for future breeding efforts.

INTRODUCTION

Alfalfa (*Medicago sativa*), the oldest plant that has been exclusively grown for forage, is the most important forage legume in the world (Quiros and Bauchan, 1988; Michaud et al., 1988). Cultivated alfalfa is a part of taxonomic group known as the *Medicago sativa-falcata* complex. There are several subspecies considered as a part of complex both in diploid (2n=2x =16) and tetraploid (2n=4x=32) levels. Hybridization is common among taxa both within the same ploidy level or across different ploidy through unreduced eggs (McCoy and Bingham, 1988). Conventionally, in addition to ploidy level, morphological traits such as flower color, pod shape, and pollen shape, have been used to differentiate among taxa in the complex. The three diploid subspecies are *M. sativa* subsp. *falcata* which has yellow flowers and sickle shaped pods, *M. sativa* subsp. *caerule*a which has purple flowers and coiled pods, and their natural hybrid, *M. sativa* subsp. hemicycla which possess variegated flower color and intermediate pod shape. The tetraploid subspecies include the direct analogue of diploid caerulea, *M. sativa* subsp. *sativa*, *M. sativa* subsp. *falcata* and the tetraploid hybrid *M. sativa* subsp. *varia*. (Quiros and Bauchan, 1988; Stanford et al., 1972).

In addition to its forage use, alfalfa has been proposed as a biofeedstock crop for cellulosic ethanol production (Delong et al., 1995). Alfalfa stems and leaves can be mechanically separated, and the leaves could be used as a high protein animal feed while the stems are used to produce energy (McCaslin and Miller, 2007; Lamb et al., 2007). Alfalfa also significantly reduces the need for fossil fuel based synthetic nitrogen fertilizers, which are expensive and can cause environmental problems (Patzek, 2004; Crews et al., 2004).

High yield and conversion efficiency are the two main qualifications needed for any crop to be considered a serious contestant for biofeedstock production (Ragauskas et al., 2006).

Genetic improvement for yield in alfalfa is not as high as those realized from the major grain crops (Hill et al., 1988). The most obvious way to increase yield is through population improvement via recurrent selection (Fehr, 1993), although limited selection for yield *per se* has been conducted in the past. A second method for yield increase is a semihybrid model, which proposes the evaluation of heterotic patterns between diverse alfalfa germplasm and the generation of population hybrids for sale (Brummer 1999). Early heterosis studies focused on the heterotic potential between subspecies *sativa* and falcata and positive results have been reported (Westgate, 1910; Waldron, 1920; Sriwatanapongse and Wilsie 1968). More recently, Riday and Brummer (2002a) have found that falcata × sativa hybrids express forage yield heterosis, but due to undesirable characteristics of falcata germplasm, such as early dormancy and slow regrowth, usage of sativa × falcata hybrids in breeding is not currently feasible (Riday and Brummer 2002b). In order to overcome the limitations of falcata germplasm, Şakiroğlu and Brummer (2007) investigated the heterosis potential of elite Midwestern cultivars when crossed to southwestern U.S. germplasm selected for adaptation in the Midwestern U.S. and concluded that those two germplasm sources to not express heterosis in dependable manner. Detecting and subsequently introgressing quantitative trait loci (QTL) that affect yield into modern cultivars could be the third way to increase yield.

Modification of cell wall components is essential to produce the quality and quantity of feedstock vital for effective bioethanol production (Farrokhi et al., 2006). In alfalfa, research to understand and manipulate cell wall components was initially conducted to increase forage quality and digestibility (Buxton and Russell, 1987; Albrecht et al., 1988; Reddy et al 2005). More recently with increased attention to cellulosic ethanol, the levels of cell wall constituents needed for an effective feedstock have been investigated. Because lignin reduces the effective

conversion of structural sugars into ethanol in the same the way that it inhibits forage digestibility in the rumen, reducing lignin is a primary goal for plant breeders and biotechnologists. Cellulose and hemicellulose are embedded in a matrix of lignin necessitating the pretreatment of biomass before fermentation in order to break down lignin and provide bacteria access to cellulose and hemicellulose for fermentation to ethanol. This is one of the most expensive procedures during cellulosic bioethanol production (Dien et al., 2006). Down-regulation of enzymes in lignin biosynthesis by genetic transformation have significantly lowered lignin content and increased digestibility (Baucher et al., 1999; Reddy et al, 2005; Gou et al., 2001a; Nakashima et al., 2000). A reduction of lignin could increase saccarification without pretreatment by increasing access to cellulose and hemicellulose (Chen and Dixon, 2007; Jackson et al., 2008).

Another approach toward increasing conversion efficiency could be the identification of QTL that control stem cell wall composition. Inheritance patterns of tetraploid alfalfa complicate genetic mapping efforts (Kaló et al., 2000), but diploid counterparts of cultivated alfalfa could be employed in genetic mapping to avoid the complexity. The genetic maps of diploid and tetraploid alfalfa are highly syntenic, and together with the possibility of interploidy hybridization, extrapolation of genetic studies conducted on diploid alfalfa to cultivated tetraploid alfalfa should be possible (Kaló et al., 2000). Knowledge of the extent of variability will be useful for the identification of appropriate germplasm for genetic mapping purposes.

We investigated the population structure of a wide range of diploid alfalfa germplasm using SSR markers (Şakiroğlu et al., manuscript in prep.). Results indicated that the three subspecies of diploid germplasm sampled were different. The SSR results also revealed that there were two distinct caerulea groups each corresponding to a geographical range (Northern vs.

southern caerulea) and two distinct falcata groups based on adaptation and ecogeography (lowland vs. upland ecotypes). Investigation of differences in biofuel potential among the five groups will help breeders more effectively use diploid germplasm for biofuel improvements.

Our objectives in the current study are: (i) to measure the variation for biofeedstock traits in the diploid gene pool of cultivated alfalfa, (ii) to investigate the population structure of wild alfalfa based on phenotype, and (iii) to compare the population structure based on phenotype to that detected using genotypic data.

MATERIALS AND METHODS

Plant Materials and Experimental Design

We selected 374 individuals from 120 diploid accessions of *M. sativa* from the USDA collections. The accessions were sampled from throughout the northern hemisphere to represent the natural distribution of diploid accessions from the three subspecies of *M. sativa* subsp. falcata, caerulea, and hemicycla. Each of accessions was represented by 1-4 individuals. All individual genotypes were grown and clonally propagated at the University of Georgia Crop & Soil Sciences Greenhouse and transplanted to the field. We conducted flow cytometry to confirm ploidy level and recorded morphological characters including flower color and pod shape that are used for taxonomic classification. Further details are presented in a companion paper (Şakiroğlu et al., 2009).

Field experiments were established on 10 May 2007 at the Univeristy of Georgia Plant Sciences Farm near Watkinsville, GA (33°52' N; 83°32' W) in a Cecil sandy loam (fine, kaolinitic, thermic typic kanhapludults) and on 18 June 2007 at the UGA Central Georgia Research and Education Center near Eatonton, GA (33°24' N; 83°29' W) in a Davidson loam (fine, kaolinitic, thermic rhodic kandiudults). At the Plant Sciences Farm, the experimental

design was a triple α-lattice, with each replication consisting of 24 incomplete blocks each with 16 entries. Ten check entries were planted to complete the design. Randomization for the block design was made using Alphagen software program. Each plot consisted of four clones planted 15 cm apart in a single row. Plots were separated within the row by 75 cm, and rows were spaced 75 cm apart. The trial at Eatonton included 205 entries planted in two replications (90 entries were only included in a single replication due to limited cuttings). Plots in each replication consisted of three clones spaced 15 cm apart with plots separated in rows by 45 cm, and rows spaced 75 cm apart. At each location, plants were clipped approximately one month after establishment and then allowed to regrow for the remainder of the season.

One harvest was taken for biomass yield on 17-18 September 2007 at Watkinsville and 24 September 2007 at Eatonton. The entire plot was clipped at a standard height of 5 cm, with the herbage placed in a paper bag, dried for 5 days at 60 ^{0}C, and weighed. The stems and leaves were separated to determine stem weight of each plot, and the stem proportion of the total biomass was computed. The number of surviving plants was recorded at the time of each harvest and was used to adjust yield to a four-plant plot basis. Plant regrowth was scored visually 2 weeks after harvest using a scale of 1 = no regrowth to 9 = vigorous regrowth. The second year harvest was taken between 5-9 June 2008 at Watkinsville, with an entire replication harvested on a single day, and on 11 June 2008 at Eatonton, using similar methods as the 2007 harvest. Plant survival was noted and regrowth ability was scored as described above. Before harvest, stem length was measured on the longest stem of each individual plant and averaged for each plot in the Watkinsville trial only. Stem thickness for each individual plot was measured in mm directly above the second node on each of four dried stems randomly selected from each plot; diameters

were averaged for each plot prior to analysis. Spring regrowth of each plot was measured in 29 March, 2008 in Watkinsville trial. Stem weight and stem proportion were not recorded in 2008.

Biofuel Analyses

Stems and leaves were separated and stems ground to pass 1-mm mesh screen in all four environments (Watkinsville 2007, Watkinsville 2008, Eatonton 2007, and Eatonton 2008). Each stem sample was scanned by near-infrared reflectance spectroscopy (Windham et al., 1983). We used NIRSystems 6500 scanning monochromator (NIRSystems, Silver Spring, MD) for collection of the reflectance measurements (log 1/*R*) between 1100 to 2500 nm, recorded at 4-nm intervals. A subset of 97 samples was selected for calibration of spectroscopy using chemical analyses. The 97 calibration samples were used for Neutral Detergent Fiber (NDF), Acid Detergent Fiber (ADF), and Acid Detergent Lignin (ADL). We determined NDF and ADF for the calibration set with fiber bag technology using the ANKOM Fiber Analyzer and the Daisy incubator (ANKOM Technology, Macedon, NY) following Vogel et al. (1999). The method of Van Soest et al. (1991) used to determine the Ash and ADL. The ANKOM bags containing the residual of the ADF procedure were placed in a 3 L Daisy^{II} incubator jar so that they could be covered with 72% H_2SO_4. 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 550 °C for 4 h, and the ash was weighed. Ash weights were calculated after accounting for the sample bag material. Hemicellulose was calculated as NDF-ADF; cellulose as ADF-ADL, and lignin as ADL-ash.

Data analyses

We conducted analyses of variance for each trait using all data over the two years and two locations in order to investigate interactions among genotypes, years, and locations. We fit a mixed linear model including replications and blocks as random effects, and years, locations, genotypes, and their interactions as fixed effects. We subsequently conducted analyses of variance by year over locations or by locations over years. We estimated least-squares means of all traits for each genotype and also for each accession (across genotypes of that accession). The 374 genotypes were allocated to one of five genetically distinct groups (southern caerulea, northern caerulea, hemicycla, lowland falcata, and upland falcata) based on SSR marker analysis (Şakiroğlu et al., 2009). Mean separations among the five groups were done using Fisher's protected least significant difference. Pearson's correlations were calculated between selected traits using mean values from the Watkinsville trial. Eatonton data were not included because only a subset of genotypes was present in that trial. All analyses were conducted using SAS 9.2 (SAS Institute, 2004). Statistical significance was assessed at the 5% probability level unless otherwise indicated.

We used 17 traits from the Watkinsville trial to compute phenotypic distances among the 374 individual genotypes. The 17 traits included the three cell wall components and TNC in each of two years and the nine agronomic traits described above. Phenotypic data were standardized and then used to calculate a Euclidian distance between all pairs of genotypes. A principal components analysis (PCA) was conducted to reduce the dimensionality of the data and enable visualization of the relationships among genotypes. We also estimated phenotypic distance and PCA of the 120 accessions in a similar manner. These analyses were conducted using NTSYS-pc (Rohlf, 1994).

We previously estimated genetic distances among genotypes and among accessions based on 89 SSR loci (Şakiroğlu et al., 2009). We then computed a correlation between the genetic and phenotypic similarity matrices for genotypes and for accessions using Mantel's test (Mantel, 1967) with the software program GenAlEx (Peakall and Smouse, 2001). Neighbor joining clustering was conducted based on the phenotypic distance matrix (NTSYS-pc; Rohlf, 1994) of 120 accessions and presented in the form of a dendogram using Dendroscope (Huson et al., 2007).

RESULTS AND DISCUSSION

Interactions of environments and locations with genotype

Hemicellulose, cellulose, lignin, TNC, dry matter yield, and regrowth following harvest were measured in each of 2 years at two locations. For these traits, two and three way interactions among genotypes, years, and locations were present for all traits, with the exception of year × location interactions for hemicellulose, cellulose, and yield (data not shown). We subsequently analyzed the data by location and by year; the year × genotype or location × genotype interactions, respectively, were significant in all cases, except for regrowth at Eatonton.

In most previous analyses, relatively little genotype × environment interaction has been observed for composition traits in alfalfa (Sheaffer et al., 2000; Sheaffer et al., 1998), even in an analysis of a large germplasm collection (Jung et al., 1997). Further, harvest timing × genotype interactions are also typically absent (Sheaffer et al., 2000; Sheaffer et al., 1998). However, in our experiment, the two harvests, one in each year, were taken at considerably different times of the year and at different developmental stages of the plant – autumn of the establishment year (2007) and early summer of the first full production year (2008). Thus, the presence of genotype × year interactions is not unexpected.

We conducted Spearman's rank correlation test for traits measured in two years in order to investigate the nature of the change between the two years. Rank correlations were highly significant ($P<0.01$) between the two years for all cell wall components and agronomic traits. The rank correlation was 0.58 for regrowth and 0.65 for yield. The rank correlations were lower for cell components, at 0.50 for hemicellulose, 0.38 for cellulose, 0.39 for lignin, and 0.36 TNC between the two years. These results indicate that ranking based on genotype performance is moderately conserved across years, but the low rank correlation implies difference of relative performance across years beyond the difference in magnitudes for the various traits. The effect of plant maturity on forage quality has been documented before (Kalu and Fick, 1983; Fick and Janson, 1990). In 2008, most genotypes had fully flowered at the time we harvested in early June, but in 2007, since harvest was in autumn, a wide variation among genotypes in maturity was evident. This could have accounted for some of the genotype × year interaction in this study. The second possible reason is the age of plants. Since 2007 was the establishment year for the plots, the cell wall composition could change in the succeeding year, as was observed by Sheaffer et al. (1998). Because of the extensive interactions, we analyzed the data separately for each year-location combination, which we denote as "environments."

Phenotypic variation and correlations among traits

Mean values for cell wall components were higher and TNC was lower in 2008 than 2007 both in Watkinsville and Eatonton (Table 3.1). Total biomass yield and regrowth was much higher in 2008 compared to 2007 reflecting the plants' development. Stem weights and the ratio of stem weights and total biomass yield was higher in Watkinsville than Eatonton in 2007 but similar in 2008; stem thickness at both locations was very similar (Table 3.2).

Wide variation among genotypes for all the traits evaluated was detected. Lignin limits the effective saccarification of cell wall sugars (Chen and Dixon, 2007), and higher lignin also results in lower forage nutritive value (Buxton and Russell, 1987; Albrecht et al., 1988; Reddy et al 2005); decreasing lignin is therefore desirable. We found that the lignin level in the unimproved diploid alfalfa stems ranged between 48 and 136 g kg^{-1} depending on the environment, or 53 - 129% of the mean. This range is comparable to that reported for transgenic reduction by down-regulation of enzymes involved in lignin biosynthesis in tetraploid alfalfa (Guo et al., 2001b; Chen and Dixon, 2007). A similar extent of variation exists for other cell wall components and agronomic traits as well, implying a great natural variability among diploid alfalfa. We compared the variation of cell wall components found in our study to that previously reported in the tetraploid alfalfa core collection (Jung et al., 1997). We used variation in three cell wall related measurements (NDF, ADF, and ADL) in 2007 and 2008 at Watkinsville, since it had the complete set of accessions. Jung et al. (1997) found that the range in NDF, ADF, and ADL was 13%, 15%, and 26% of mean, respectively. We found that the range of variation in NDF was 52% and 40% of mean in 2007 and 2008, respectively. The range of ADF was 67% and 48%, respectively, and the variation of ADL was 71% and 50%, respectively. Therefore, thed variation detected in diploid germplasm was 2-4 fold higher compare to those reported in the tetraploid core collection. The variation in these traits could be used by selecting extreme phenotypes for mapping the genetic control of forage quality, biofuel potential, and yield in diploid accessions, thereby avoiding complications arising from tetrasomic inheritance. Our analysis of individual genotypes enables us to accurately select parents for future screening.

The Northern caerulea accessions PI577545 and PI577551 and upland falcata accessions PI631817 and SD201 were high in cellulose and hemicellulose content and also had consistently

higher yield in both years (data not shown). Due to high correlation among cell wall components, high yielding accessions with higher hemicellulose and cellulose content were also high in terms of lignin content. The two accessions from the upland falcata group, PI631556 and PI631707, had high cellulose content but considerably low lignin content. The former was high in total biomass yield, but the latter accession had yield near the average of all accessions.

Based on mean values of individual genotypes at grown at Watkinsville, phenotypic correlations among cell wall constituents in the same year were high (Table 3.3). Correlation of the same traits in two different years for the cell wall components were around 0.50. The correlation of cell wall components and agronomic traits were generally higher if the trait data were gathered in the same year, except for the correlation between total biomass yield in 2008 and hemicellulose in 2007, which showed a somewhat stronger correlation (0.31) than the other correlations between these traits. Total biomass yield was correlated between years ($r = 0.64$), as was regrowth ability (0.58). Correlation coefficients between cellulose and lignin in both years were positive and very high (0.70). Lignin in 2007 did not have significant correlation with total biomass yield 2007 (Table 3.3).

Genotype and accession clustering

Our previous molecular marker analysis indicated that diploid germplasm can be differentiated into five main groups, corresponding to northern and southern caerulea, upland and lowland falcata, and hemicycla (Şakiroğlu et al., 2009). Therefore, we evaluated differences in phenotypic traits among these five groups. The groups differed for all traits (Table 3.4). From a bioenergy perspective, the northern caerulea group has most value, having the highest or among the highest values for lignin and cellulose concentrations in 2008, biomass yield in both years, plant height, spring recovery, and regrowth after harvest (Table 3.4). Upland falcata accessions

also show promise by having high cellulose, hemicellulose, and yield, although in general, they have slightly less overall value than northern caerulea. The yield of southern caerulea, lowland falcata, and hemicycla were considerably lower than the others, limiting their value for biofuel uses. Southern caerulea had the highest concentrations of cellulose and lignin in the establishment year, and may be a source of alleles for increased concentrations of these compounds.

We conducted a principal components analysis to reduce the dimensionality of the phenotypic data and enable us to observe relationships among entries. The first two principal components accounted for 56% of the variation among genotypes (39% for PC1 and 17% for PC2). Plotting the first two principal components did not show any clear clustering of genotypes (Figure 3.1). We were interested to determine if phenotypic differences would reflect the differentiation we had observed using genetic markers. Our evaluation of population structure within diploid *M. sativa* germplasm clearly separated falcata from caerulea, leaving hemicycla in between, and provided strong evidence for further hierarchical separation of both falcata and caerulea into two groups (Şakiroğlu et al., 2009). The phenotypic data did not reflect that separation. A Mantel test comparing distance matrices based on genotypic and phenotypic data showed that the two matrices were different ($P < 0.001$), as expected based on the visual clustering.

A distance matrix computed based on phenotypic data was used to develop a neighbor joining dendrogram showing relationships among accessions, rather than individual genotypes. Three clusters of accessions were evident (Figure 3.2), but these do not correspond to those found with the molecular marker analysis. Each of the three clusters included accessions from all five molecular-marker defined groups. Cellulose, hemicellulose, and lignin content in both 2007

and 2008 were higher and TNC was lower in clusters 1 and 2 compared to cluster 3 (data not shown). For yield and agronomic traits, cluster 1 outperformed cluster 2 in all nine traits (Figures 3.2 and 3.3).

We conducted PCA evaluating accessions (Fig. 3.3). The first principal component captured variation in yield and related agronomic traits, and accounted for 42% of the variation in the population. The second principal component, accounting for 23% of the variation, was associated with stem composition. Plotting the first two principal components derived from an analysis of the phenotypic data of the 118 accessions did not cluster accessions into distinct subgroups (Figure 3.1), in accordance with the phenotypic analysis of the individual genotypes.

Practical implications

The lack of a relationship between the molecular marker based clustering and phenotypic clustering implies that desirable alleles for improvement of yield and cell wall traits (either in the context of forage quality or biofuel potential) are dispersed across all subspecies. Certain germplasm sources – the northern caerulea accessions, for example – offer the best starting point for accessing new genes for cultivated alfalfa. Nevertheless, other germplasm, particularly certain falcata accessions, should be carefully evaluated for the traits of interest as these may contain novel alleles, as suggested by their molecular marker distinctiveness. The next step in using this material is to develop populations, select at the diploid level for agronomic utility, and then scale to the tetraploid level using unreduced gametes, following an analytic breeding scheme (Chase, 1963), or directly tetraploidize individual plants and use them as donors of specific alleles based on future mapping efforts. A key finding in this experiment is that diploid germplasm has wide variation, and could serve as a useful reservoir of beneficial alleles for cultivated tetraploid alfalfa.

REFERENCES

Albrecht, K. A., W. F. Wedin, and D. R. Buxton. 1987. Cell-wall composition and digestibility of alfalfa stems and leaves. *Crop Sci.* 27**,** 735–741.

Baucher, M., M.A. BernardVailhe, B. Chabbert, J.M. Besle, C. Opsomer, M. VanMontagu, and J.Botterman. 1999. Down-regulation of cinnamyl alcohol dehydrogenase in transgenic alfalfa (*Medicago sativa* L.) and the effect on lignin composition and digestibility. *Plant Molecular Biology* 39: 437–447.

Brummer, E.C. 1999. Capturing heterosis in forage crop cultivar development. Crop Sci. 39:943–954.

Buxton, D. R. and J. R. Russell. 1988. Lignin constituents and cell-wall digestibility of grass and legume stems. *Crop Sci.* 28, 553–558.

Chen, F., and R.A. Dixon. 2007. Lignin modification improves fermentable sugar yields for biofuel production. Nature Biotechnology 25, 759-761.

Crews, T.E. and M.B. Peoples. 2004. Legume versus fertilizer sources of nitrogen: Ecological tradeoffs and human needs. Agriculture, Ecosystems and Environment 102:279-297.

Delong, M.M., D.R. Swanberg, E.A. Oelke, C. Hanson, M. Onischak, M.R. Schmid, and B.C. Wiant. 1995. Sustainable biomass energy production and rural economic development using alfalfa as a feedstock. p. 1582–1591. *In* D.L. Klass (ed.) Second Biomass Conf. of the Americas: Energy, Environment, Agriculture, and Industry, Portland, OR. 21–24 Aug .

Dien, B.S., H-J.G. Jung, K.P. Vogel et al. 2006. Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. Biomass Bioenergy 30:880–891.

Farrokhi, N., R.A. Burton, L. Brownfield, M. Hrmova, S.M. Wilson, A. Bacic, G.B. Fincher. 2006. Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. Plant Biotechnology Journal 4:145–167.

Fehr, W.R. 1987. Principles of cultivar development. Macmillan, NY.

Fick, G.W., and C.G. Janson. 1990. Testing mean stage as a predictor of alfalfa forage quality with growth chamber trials. Crop Sci. 30:678–682.

Griffin, T.S., K.A. Cassida, O.B. Hesterman, and S.R. Rust. 1994. Alfalfa maturity and cultivar effects on chemical and in situ estimates of protein degradability. Crop Sci. 34:1654–1661.

Guo, D., F. Chen, K. Inoue et al. 2001a. Down-regulation of Caffeic Acid 3-*O*-methyltransferase and Caffeoyl CoA 3-*O*-methyltransferase in transgenic alfalfa (*Medicago sativa* L.): Impacts on lignin structure and implications for the biosynthesis of G and S lignin. Plant Cell 13:73–88.

Guo D., F. Chen., J. Wheeler et al. 2001b Improvement of In-rumen Digestibility of Alfalfa Forage by Genetic Manipulation of Lignin *O*-methyltransferases. Transgenic Res 10:457–464.

Hill, R.R., Jr., J.S. Shenk, and R.F Barnes. 1988. Breeding for yield and quality. p. 809–825. *In* A.A. Hanson et al. (ed.) Alfalfa and alfalfa improvement. ASA–CSSA–SSSA, Madison, WI.

Huson, D.H., D.C. Richter, C. Rausch, T. Dezuian, M. Franz, and R. Rupp. 2007. Dendroscope: an interactive viewer for large phylogenetic trees, *BMC Bioinformatics* 8**:** 460.

Jackson, L.A., G.L. Shadle, R. Zhou, J. Nakashima, F. Chen, and R.A. Dixon. 2008. Improving saccharification efficiency of alfalfa stems through modification of the terminal stages of monolignol biosynthesis. Bioenergy Research 1, 180-192.

Jung, H.G., C.C. Sheaffer, D.K. Barnes, and J.L. Halgerson. 1997. Forage quality variation in the U.S. alfalfa core collection. Crop Sci. 37:1361-1366.

Kalu, B.A., and G.W. Fick. 1981. Quantifying morphological development of alfalfa for studies of herbage quality. Crop Sci. 21:267–271.

Lamb, J.F.S., H.G. Jung, C.C. Sheaffer, D.A. Samac. 2007. Alfalfa leaf protein and stem cell wall polysaccharide yields under hay and biomass management systems. *Crop Science*, 47:1407-1415.

Mantel, N. 1967. The detection of disease clustering and a generalized Multi- regression approach. Cancer Res. 27:209–220.

McCaslin, M. and D. Miller. 2007. The future of alfalfa as a biofuels feedstock. 37th California Alfalfa & Forage Symposium. December 17-19, 2007 Monterey, CA.

McCoy, T.J. and Bingham E.T. 1988. Cytology and cytogenetics of alfalfa. p. 739-776, *In* A.A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Michaud, R., W.F. Lehman and M.D. Rumbaugh. 1988. World distribution and historical development p. 26-82, *In* A.A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Nakashima, J., F. Chen, L. Jackson et al. 2008. Multi-site genetic modification of monolignol biosynthesis in alfalfa (*Medicago sativa* L.): effects on lignin composition in specific cell types. New Phytol 179:738–750.

Peakall, R., and P.E. Smouse. 2001. GenAlEx V5; Genetic analysis in excel. Population genetic software for teaching and research. Australian National University, Canberra, Australia. http://www.anu.edu.au/BoZo/GenAlEx.

Quiros, C.F., and G.R. Bauchan. 1988. The genus *Medicago* and the origin of the *Medicago sativa* complex, p. 93-124, *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Ragauskas, A.J., C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, Jr., J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer, and T. Tschaplinski. 2006. The path forward for biofuels and biomaterials. Science 311:484-489.

Reddy, M.S.S., F. Chen, G. Shadle, L. Jackson, H. Aljoe, and R.A. Dixon. 2005. Targeted down-regulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (*Medicago sativa* L.). Proc Natl Acad Sci USA 102:16573–16578.

Riday, H. and E.C. Brummer. 2002a. Forage yield heterosis in alfalfa. Crop Sci 42:716–723

Riday, H. and E.C. Brummer. 2002b. Heterosis of agronomic traits in alfalfa. Crop Sci.

Rohlf, F.J. 1994. NTSYS-pc: Numerical taxonomy and multivariate analysis system, Ver. 1.80. Exeter Software, Setauket, NY42:1081–1087.

Şakiroğlu, M. and E.C. Brummer. 2007. Little Heterosis between Alfalfa Populations Derived from the Midwestern and Southwestern United States. Crop Sci 47:2364-2371

SAS Institute Inc., SAS® 9.1.2Cary, NC: SAS Institute Inc., 2004.

Sheaffer, C.C., D. Cash, N.J. Ehlke, J.C. Henning, J.G. Jewett, K.D. Johnson, M.A. Peterson, M. Smith, J.L. Hansen, and D.R. Viands. 1998. Entry x environment interaction for alfalfa forage quality. Agron. J. 90:774–780.

Sheaffer, C.C., N.P. Martin, J.F.S. Lamb, G.R. Cuomo, J.G. Jewett, and S.R. Quering. 2000. Leaf and stem properties of alfalfa entries. Agron J 92:733-739.

Sriwatanapongse, S., and C.P. Wilsie. 1968. Intra- and intervariety crosses of *Medicago sativa* L. and Medicago falcata L. Crop Sci. 8:465–466.

Stanford, E.H., W.M. Clement, and E.T. Bingham. 1972. Cytology and evolution of the Medicago sativa-falcata complex, p. 87-100, *In* C.H. Hanson(ed.) Alfalfa science and technology. ASA–CSSA–SSSA, Madison, WI.

Van Soest, P.J., J.B. Robertson, and B.A. Lewis. 1991. Symposium: carbohydrate methodology, metabolism, and nutritional implications in dairy cattle. J. Dairy Sci. 74:3583–3597.

Vogel, K.P., J.F. Pedersen, S.D. Masterson, and J.J. Troy. 1999. Evaluation of a filter bag system for NDF, ADF, and IVDMD forage analysis. Crop Sci. 39 : 276–279

Westgate, J.M. 1910. Variegated alfalfa. USDA Bur. PI Ind. Bull. 169:1–63.

Waldron, L.R. 1920. First generation crosses between two alfalfa species. J. Am. Soc. Agron. 12:133–143.

Windham, W.R., D.R. Mertens, and F.E. Barton II. 1989. Protocol for NIRS calibration: sample selection and equation development and validation. e.c. 96–103. *In* G.C. Marten et al. (ed.) Near infrared reflectance spectroscopy (NIRS): Analysis of forage quality. USDA Agric. Handbook 643.

Table 3.1. Mean, standard deviation, and range (in parenthesis) of cell wall components of 372 wild diploid alfalfa genotypes over two Georgia locations (Athens and Eatonton) and over two years (2007 and 2008).

	Hemicellulose	Cellulose	Lignin	TNC
	g/kg	g/kg	g/kg	g/kg
Athens				
2007	150±5	290±2	90±5	170±119
	(130-174)	(183-388)	(48-111)	(109-294)
2008	160±5	380±18	110±5	100±10
	(129-197)	(259-464)	(78-136)	(69-166)
Overall	150±5	340±17	100±5	130±15
	(129-171)	(241-399)	(67-117)	(96-226)
Eatonton				
2007	150±5	290±7	90±3	140±12
	(132-171)	(222-350)	(70-116)	(102-204)
2008	160±5	360±2	110±5	120±10
	(138-185)	(232-458)	(70-130)	(75-205)
Overall	160±5	330±18	100±5	130±13
	(143-173)	(262-391)	(86-120)	(81-192)
	(137-180)	(228-419)	(71-129)	(92-192)
2007	150±5	290±16	90±5	160±19
	(128-177)	(183-355)	(49-112)	(106-293)
2008	150±5	380±18	110±5	110±10
	(131-188)	(282-459)	(85-135)	(75-159)
Overall	150±5	340±18	100±5	130±15
	(130-169)	(240-400)	(61-123)	(97-226)

Table 3.2. Mean standard deviation and range (in parenthesis) of agronomic traits of 372 wild diploid alfalfa genotypes over two Georgia locations (Athens and Eatonton) and over two years (2007 and 2008).

	Dry matter yield	Regrowth	Stem total mass ratio	Stem weight	Spring regrowth	Stem length	Stem thickness
	g plot^{-1}	0-9		g plot^{-1}	0-9	cm	mm
Athens							
2007	59 ± 33	2.7±0.8	0.41±0.1	24±17			
	(0.25-217)	(0.8-5.0)	(0.1-0.9)	(0.6-98)			
2008	381±210	3.5±0.9			2.3±0.7	81±14	2.3±0.4
	(7.4-2247)	(0.5-6.7)			(0.3-6.4)	(26-133)	(0.5-4.2)
Overall	220±152	3.1±0.9					
	(3-1226)	(0.9-5.2)					
Eatonton							
2007	43±19	3.2±1	0.33±0.16	14±9			
	(1-202)	(0.7-9.8)	(0.1-0.6)	(1-68)			
2008	399±210	2.8±0.8					2.3±0.5
	(2-1672)	(0.2-5.8)					(0.7-6.0)
Overall	215±141	3±0.9					
	(0-855)	(0-7.8)					
	(10-854)	(0-7.5)					
2007	55±32	2.8±0.-8	0.4±0.1	23±16			
	(5-226)	(1.3--4.8)	(0.1-0.7)	(0.8-104)			
2008	384±209	3.4±0.9					2.3±0.4
	(20-2153)	(0.5-6.5)					(0.4-4.2)
Overall	219±151	3.1±0.9					
	(14-1183)	(1.2-5.3)					

Table 3.3. Pearson's correlation coefficients among selected cell wall constituents and agronomic traits.

Trait[†]	Hemi 2007	Cell 2007	Lignin 2007	TNC 2007	Hemi 2008	Cell 2008	Lignin 2008	TNC 2008	Yield 2007	Regr 2007	Height 2008	Thick 2008	Yield 2008
Cell - 2007	0.58[‡]												
Lignin - 2007	0.73	0.70											
TNC - 2007	-0.52	-0.75	-0.75										
Hemi - 2008	0.50	0.38	0.31	-0.26									
Cell - 2008	0.26	0.45	0.34	-0.41	0.33								
Lignin - 2008	0.30	0.24	0.50	-0.46	0.22	0.70							
TNC - 2008	-0.21	-0.18	-0.36	0.49	-	-0.69	-0.83						
Yield - 2007	-	0.33	0.16	-0.17	0.14	-	-	-					
Regr - 2007	-	0.16	-	-	-	-	-0.22	0.22	0.59				
Height - 2008	0.39	0.57	0.47	-0.43	0.16	0.32	0.24	-0.21	0.34	0.16			
Thick - 2008	0.52	0.68	0.57	-0.50	0.27	0.30	0.20	-0.18	0.39	0.22	0.67		
Yield - 2008	0.31	0.53	0.41	-0.34	0.17	0.21	-	-	0.64	0.44	0.59	0.66	
Regr - 2008	-	0.26	-	-	-	-	-	0.14	0.46	0.58	0.32	0.27	0.47

[†] Hemi = Hemicellulose, Cell = Cellulose, TNC = Total Non-structural Carbohydrates, Regr = Regrowth after harvest.
[‡] Correlations shown are significant at the 1% probability level; Non-significant correlations are designated with "-".

Table 3.4. Mean of cell wall components and agronomic traits of the five main population of diploid alfalfa in the Athens trial over two years (2007 and 2008).

		Population				
Trait	Units	Southern caerulea	Northern caerulea	Hemicycla	Lowland falcata	Upland falcata
		2007				
Hemicellulose	g kg^{-1}	143e[†]	147c	149b	144d	151a
Cellulose	g kg^{-1}	302a	298b	291c	272d	287c
Lignin	g kg^{-1}	92a	87c	90b	78e	83d
TNC[‡]	g kg^{-1}	144d	160b	152c	184a	182a
Biomass yield	g plot^{-1}	44c	77a	34d	60b	66b
Stem proportion		0.43b	0.45a	0.43a,b	0.34d	0.40c
Regrowth	score	2.4d	2.9a	2.6c	3.1a	2.7b
		2008				
Hemicellulose	g kg^{-1}	154c	159b	160a	152d	159a,b
Cellulose	g kg^{-1}	365c	396a	381b	357d	396a
Lignin	g kg^{-1}	113c	118a	117a,b	106d	117b
TNC	g kg^{-1}	105b	97d	97d	112a	101c
Biomass yield	g plot^{-1}	294c	632a	245d	254c,d	439b
Spring growth	score	2.2c	2.8a	1.9d	1.7d	2.6b
Regrowth	score	3.5b	3.5a,b	2.1c	3.7a	3.7a
Plant height	cm	81b	92a	76c	68d	83b
Stem thickness	mm	2.07c	2.89a	2.11c	1.72d	2.51b

[†]Values within rows followed by different letters are significantly different at the 5% probability level.
[‡]TNC = Total Nonstructural Carbohydrates.

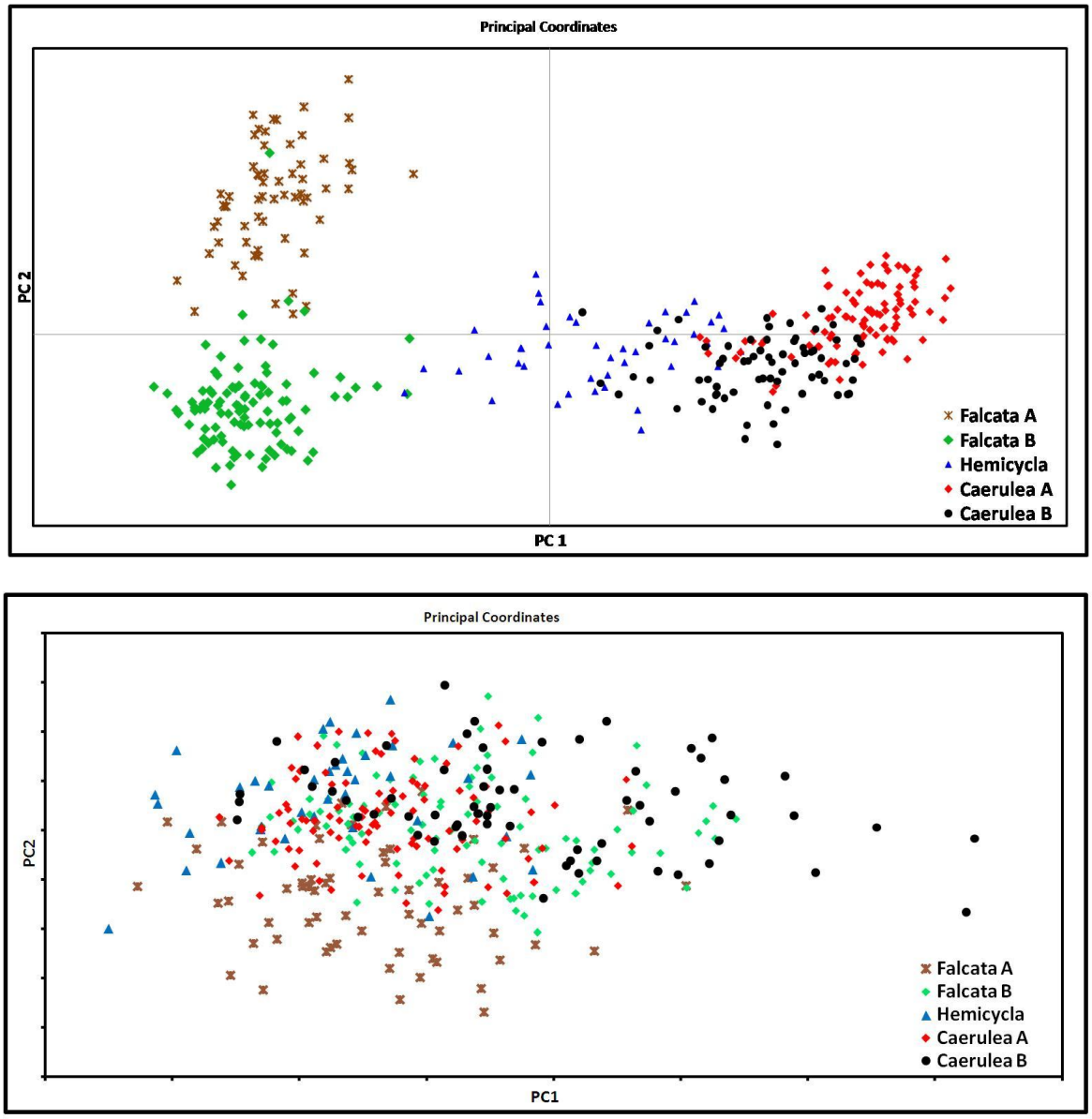

Figure 3.1. Principal components analysis of 374 diploid alfalfa genotypes from five populations based on (A) 89 polymorphic SSR markers (B) 17 phenotypic traits measured in two years at Watkinsville, GA.

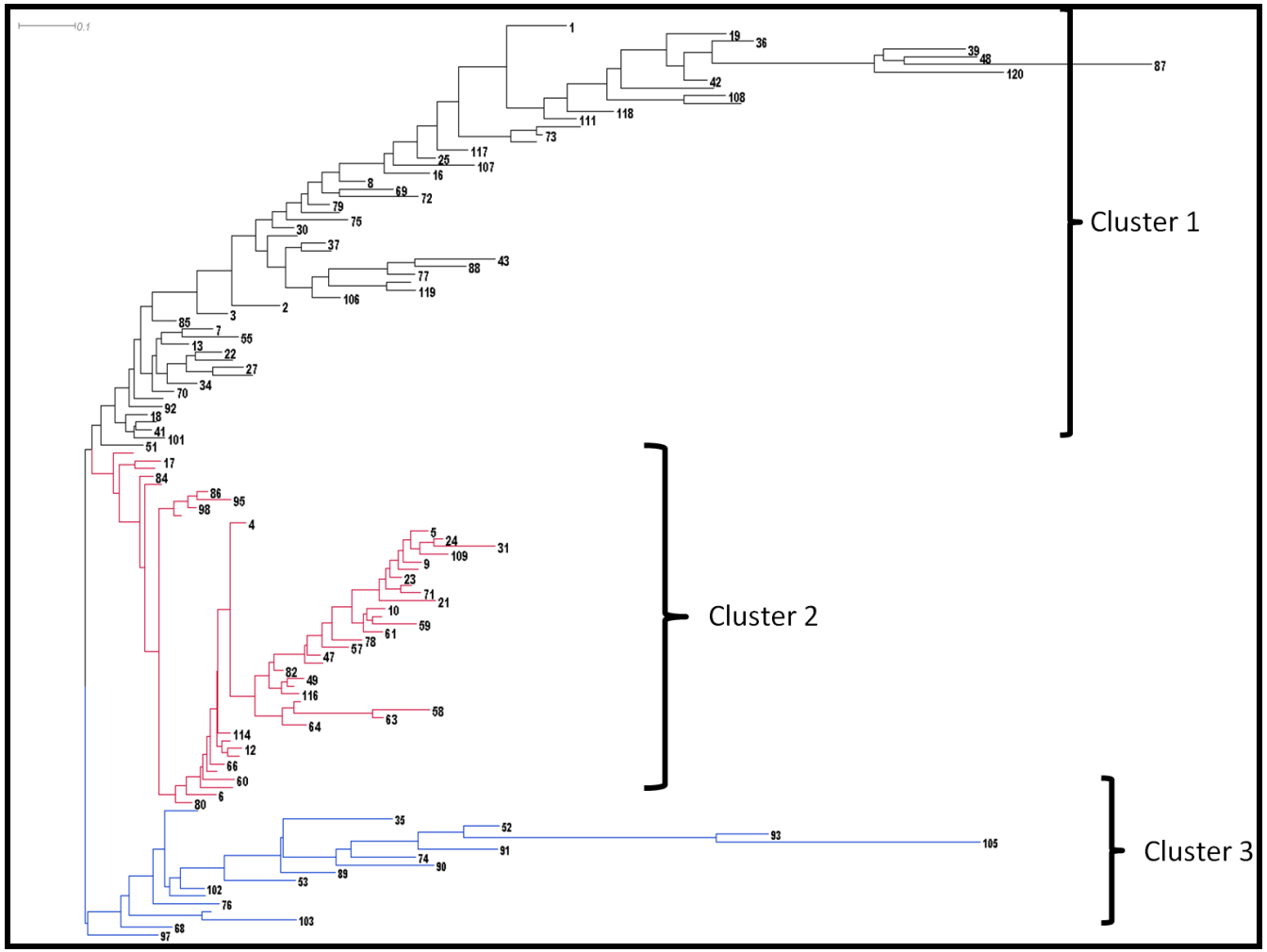

Figure 3.2. Neighbor-Joining dendogram of 120 diploid alfalfa accessions based on 17 cell wall and agronomic traits measured in 2007 and 2008 at Watkinsville, GA. The three different colors represent different clusters.

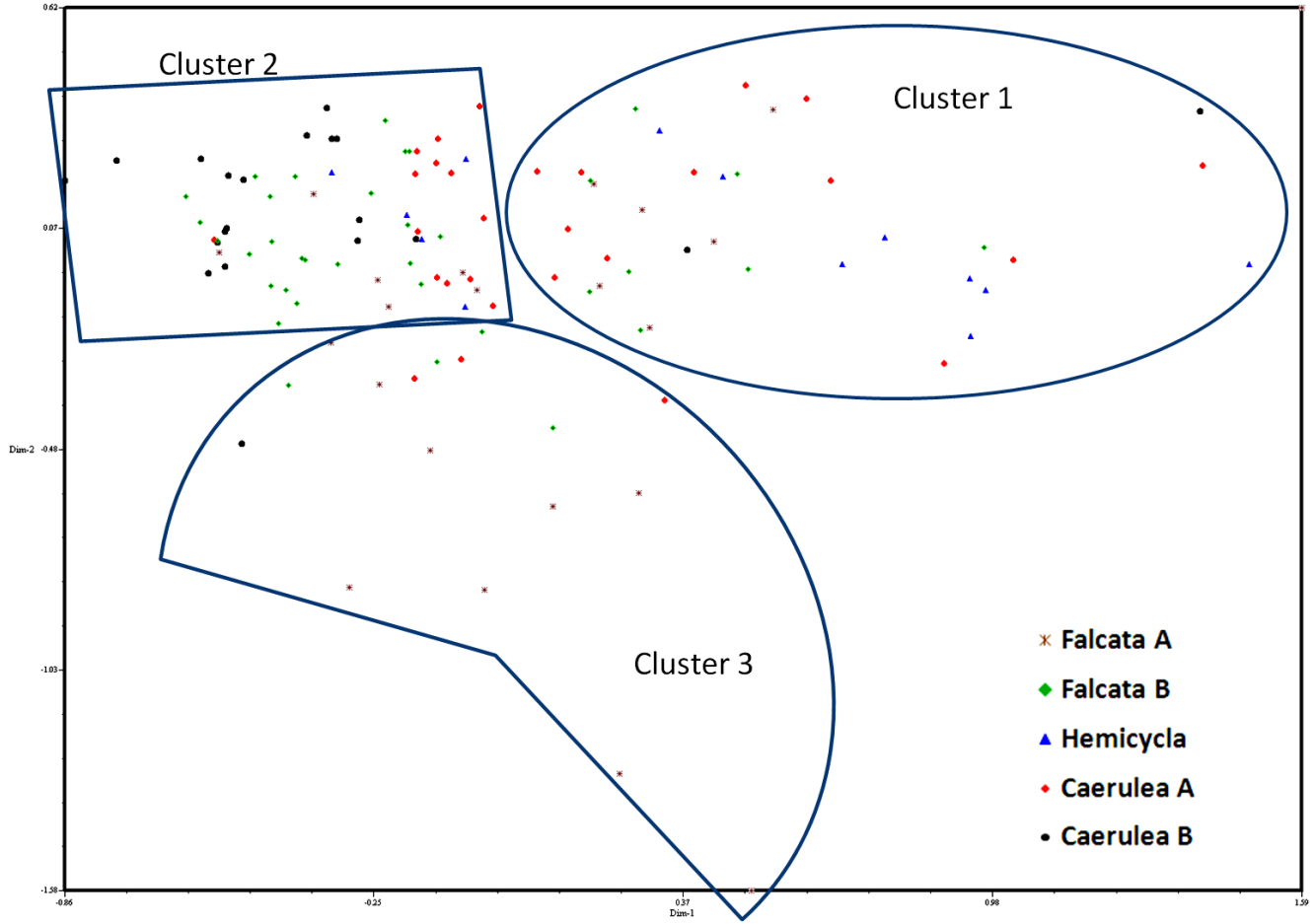

Figure 3.3. Principal components analysis of 120 diploid alfalfa accessions measured for 17 cell wall and agronomic traits in 2007 and 2008 at Watkinsville, GA. The plot is based on the first two principal components.

CHAPTER 4

PATTERNS OF LINKAGE DISEQUILIBRIUM AND ASSOCIATION MAPPING IN DIPLOID ALFALFA (*M. SATIVA* L.)[3]

[3] Sakiroglu, M., S. Sherman-Broyles, J.J. Doyle, K. J. Moore, and E. C. Brummer. To be submitted to *Crop Science*

ABSTRACT

Association mapping enables the detection of marker-trait associations in unstructured populations by taking advantage of relic linkage disequilibrium (LD) that exists between markers and the actual locus affecting the trait. Our objective was to understand the pattern of LD decay in the diploid alfalfa (*Medicago sativa* L.) genome so that we could assess the potential of using association mapping to identify markers linked to biomass yield and cell wall composition traits. We scored 89 highly polymorphic SSR loci on 374 unimproved diploid alfalfa genotypes from 120 accessions to infer genomewide patterns of LD. We also sequenced an ~500bp fragment of a lignin biosynthesis gene (F5H) to identify single nucleotide polymorphisms (SNP) and infer within gene estimates of LD. We conducted association mapping for cell wall components and agronomic traits using the SSR markers and the F5H gene sequence SNP. We found extensive LD among SSR markers, extending over 10 Mb. In contrast, within gene LD extended about 200bp and sharply declined for longer distances. We identified one SSR and one SNP associated with biomass yield, but no other associations were detected. Based on our results, focusing association mapping on candidate gene sequences will be necessary until a dense set of genomewide markers is available for alfalfa.

INTRODUCTION

Linking DNA polymorphism to trait phenotypic variation is an increasingly important tool for plant breeding programs (Lande and Thompson, 1990). Historically, segregating populations of a particular cross have been used to identify marker-trait associations (Stuber et al., 1999). More recently, association mapping has shown promise due to the increased access to abundant molecular markers in many crops (Stich et al., 2005).

Association mapping takes advantage of recombination that occurred over many generations of historical matings and which has decreased linkage disequilibrium (LD) to short chromosomal intervals. This enables statistically robust marker-trait associations to be detected (Jannink and Walsh, 2002). Association mapping can be conducted within existing breeding populations, accounting for all allele variation in the population unlike typical biparental mapping populations (Hirschhorn and Daly, 2005; Remington et al., 2001). In general, the precision of locating a QTL is much higher in association panels compared to biparental populations provided sufficient markers are available to detect the QTL. However, if LD extends only over short distances, the biparental mapping approach is more powerful to detect the existence of a QTL, particularly if marker numbers are limited (Mackay and Powell, 2007).

Two major drawbacks exist in association mapping. First, false positive associations between markers and traits can be detected due to the presence of population structure (Aranza et al., 2005; Lander and Schork, 1994). To avoid this problem, population structure can be assessed with marker information from genome wide genetic markers (such as SSRs), and association tests can then be conditioned on the population structure to reduce the false positive rate (Aranza et al., 2005; Pritchard et al., 2000). Second, the extent of LD plays a practical role in determining the number of markers needed to detect associations between genotype and phenotype. If the rate

of LD decay is high, the association between alleles at different loci will disappear quickly. In such a situation, detection of association between genetic markers and phenotype of interest requires more markers (Hagenblad and Nordborg, 2002). When the limiting factor for association mapping is the absence of a large number of markers evenly dispersed throughout the genome, another strategy that can be used for detection of marker-trait associations is to assay variation in candidate genes (Neale and Savolainen, 2004). Tracking genotypic variants via single nucleotide polymorphism (SNP) markers in candidate genes and associating this polymorphism with phenotypic variation may be a better approach than using a low density of randomly chosen markers scattered throughout the genome. For both cases, in order to more effectively design and use association studies, knowledge of the LD structure in the genome is needed (Oraguzie et al., 2007).

Alfalfa is one of the most important forage legumes in the world (Quiros and Bauchan, 1988; Michaud et al. 1988). Alfalfa has the potential to reduce contemporary environmental problems arising from agricultural practices because it's perennial nature limits soil erosion and provides wildlife habitat, and its nitrogen fixation ability as a legume eliminates the need for synthetic nitrogen application (Patzek, 2004; Crews et al., 2004), which could be eliminated or greatly reduced by the cultivation of alfalfa. In addition, alfalfa has recently been proposed as a bioenergy crop (Delong et al., 1995). Alfalfa stems and leaves can be separated and the leaves could be used as a high protein animal feed while the stems would be used to produce energy (McCaslin and Miller, 2007; Lamb et al., 2007).

Higher total biomass yield and better conversion efficiency are crucial for any biofeedstock crop candidate (Ragauskas et al., 2006). Genetic improvement for yield in alfalfa is not as high as those realized for the major grain crops (Hill et al 1988). Identification of QTL

that are associated with yield and incorporation of them into modern cultivars could enhance the efficiency of alfalfa breeding to overcome the current yield stagnation problem. The other main challenge of robust cellulosic bioethanol production is the effective hydrolysis of cellulose and solubilization of hemicellulose in the presence of lignin (U.S. DOE., 2006). Reducing lignin content can increase the efficiency of sugar release from cell wall complexes up to two fold (Chen and Dixon, 2007). Mapping QTL for lower lignin and elevated cellulose content could lead to the development of superior bioenergy varieties.

Cultivated alfalfa is an autotetraploid and inheritance patterns complicate genetic mapping. The diploid counterparts of cultivated alfalfa could be employed in genetic mapping to avoid the complexity. The genetic maps of diploid and tetraploid alfalfa are highly syntenic, and together with the possibility of hybridization across the two ploidy levels, extrapolation of genetic studies conducted on diploid alfalfa to cultivated tetraploid alfalfa should be possible (Kaló et al., 2000).

The extent of LD in alfalfa is unknown. It is vital to understand the patterns of LD across the entire genome and within genes both within and among populations to determine strategies for mapping underlying genes (Rafalski and Morgante, 2004). In this paper, we estimated the extent of LD among 89 polymorphic SSR loci distributed throughout genome and in the ~500bp transcribed region of the *ferulate 5-hydroxylase (F5H)* gene that is known to be involved in lignin biosynthesis using 374 unimproved diploid alfalfa genotypes from 120 accessions. We also evaluated sequence polymorphism in the *F5H* gene. We used association mapping to detect relationships between SSR and SNP marker polymorphisms and 17 traits.

MATERIALS AND METHODS

Plant materials and phenotyping

We selected 374 individual genotypes from 120 accessions obtained from the USDA National Plant Germplasm System, representing the geographical distribution of the diploid *M. sativa* complex, including subsp. *caerulea*, *falcata*, and *hemicycla*. Further information on these genotypes can be found in Şakiroğlu et al. (2009a). These genotypes were planted in field experiments near Watkinsville and Eatonton, Georgia. The experimental design and procedures were reported previously (Şakiroğlu et al., 2009b). We evaluated stem hemicellulose, cellulose, lignin, and total nonstructural carbohydrate (TNC) composition, total aboveground biomass yield, and regrowth after harvest in 2007 and 2008. Five other agronomic traits were measured in one year, stem yield and stem/leaf ratio in 2007, and plant height, stem thickness, and spring regrowth in 2008.

Genotyping and sequencing

We scored 89 SSR loci on the 374 genotypes and analyzed genetic relationships among them, as described previously (Şakiroğlu et al., 2009a). The physical location of the SSR markers was determined using BLAST to find the sequence of the SSR primers or the EST from which the SSR marker was developed on the genome sequence of *M. truncatula*, version 2 (www.medicago.org) (Figure 4.1 and Figure 4.2).

We evaluated the sequence of a 495bp fragment of the ferulate-5-hydroxylate (F5H) gene involved in lignin biosynthesis. The published sequence of F5H from *M. sativa*, Genbank Accession DQ222912 (Reddy *et al.* 2005), was used to query the *M. truncatula* genome sequence version 2. BLAST results indicated that the gene resides on chromosome 5 in *M.*

truncatula, and has an intron that is considerably larger than the one in *Arabidopsis thaliana* (www.tigr.org/tigr-scripts/euk_manatee/shared/ORF_infopage.cgi?db=ath1&orf=At5g04330).

Two primers, JZF5H3F and JZF5H3R, were designed in exon 1 of the F5H gene, which produce a ~500 bp fragment (Figure 4.3). PCR products were generated from genotypes using a standard PCR program of (1) 5 minutes at 94˚C, (2) 40 cycles of 94˚C for 1 minute, 58˚C annealing temperature for 1 minute and 72 ˚ C for 1 minute, and (3) 8 minute extension at 72˚C on Techne thermocyclers. The PCR included 67mM Tris HCl pH 8, 2mM $MgCl_2$, 250µM dNTPs, 2% DMSO, and 0.2µM primers. Products were visualized on 1% agarose gels stained with ethidium bromide to assess quality prior to sequencing. Direct sequencing of PCR products was conducted at the Cornell Life Sciences Core Facility on ABI3730 sequencers (Applied Biosystems Inc.). Electropherograms were examined by eye and double peaks, indicating the presence of two different bases in heterozygotes were coded using standard International Union of Biochemistry ambiguity codes. Alignments were made using default parameters of ClustalW in BioEdit (Hall, 1999) and editing by eye. Consensus sequences were constructed by contiging forward and reverse sequences using Sequencher (GeneCodes, Inc., Ann Arbor, MI).

Data analysis

We used both the genome-wide SSR marker data and the sequence data from F5H for association mapping of quantitative trait loci (QTL) for the phenotypic trait data from the field experiment. False positive marker-trait associations can arise due to population structure among the genotypes evaluated rather than from linkage. We inferred population structure using the software program STRUCTURE (Pritchard et al., 2000) as described previously (Şakiroğlu et al., 2009a). In brief, the most likely true number of subpopulations (K) was five, with each of the groups corresponding to biologically meaningful divisions. The three subspecies clearly

separated into distinct clusters, and the subspecies falcata and caerulea were each further divided into two subgroups. Hence, in the association analyses, we used K=5 for grouping (Şakiroğlu et al., 2009a). SPAGeDi 1.2 software (Hardy and Vekemans 2002) was used to estimate a kinship matrix for each pair of genotypes (Ritland, 1996) using the 89 SSR loci. Negative kinship values were set to zero, following Yu et al. (2006).

Linkage disequilibrium among SSR markers was calculated using the software program GENEPOP 4.0 (Raymond and Rousset, 1995), and LD between pairs of loci whose physical location was obtained from the *M. truncatula* genome sequence was reported. The LD between polymorphic sites in the F5H sequence was estimated using TASSEL 2.1 (Bradbury et al., 2007). The LD based on sequence data and SNPs was estimated by squared allele-frequency correlations (r^2) for pairs of loci (Hill and Robertson, 1968). Marker pairs that showed LD with a p value of 0.00001 or lower were considered to be real and highest value of -log (p-value) was set to 5. Because we used a very large number of tests, corrections for multiple testing were performed using the positive false discovery rate (FDR) method (Storey 2002; Storey and Tibshirani 2003) implemented in the software program Q-Value (Storey 2002).

In order to estimate genetic diversity in the F5H gene segment, the average pairwise difference between sequences, π, (Tajima, 1983), and Watterson's estimator of θ (Watterson, 1975) were calculated using the computer program DnaSP v4 (Rozas et al., 2003).

Least squares means of agronomic traits of the Watkinsville trial were obtained using PROC MIXED of SAS 9.2 (SAS Institute Inc., 2004). The software program TASSEL 2.1 (Bradbury et al., 2007) was used for detection of the association between SNP or SSR markers and the phenotypic data. A mixed linear model (MLM) was fitted for each single marker and trait (Yu et al. 2006). This approach takes into account relatedness among individuals by using the pairwise

kinship matrix as a covariate in addition to the population structure inference (Q matrix) in the mixed model. Correction for multiple testing was applied to P-values obtained from MLM using the positive false discovery rate (FDR) method (Storey 2002; Storey and Tibshirani 2003) implemented in software program Q-Value (Storey, 2002).

RESULTS

Diversity of sequences

The two indications of sequence variation in F5H gene, π and θ, were estimated for all five groups of wild diploid alfalfa as well as over all genotypes (Table 4.2). θ (theta) is defined as the number of polymorphic sites in a sample of DNA sequences corrected for sample size whereas, π is defined as the expected heterozygosity per nucleotide site (Hartl and Clark, 1997). Based on these estimates of variation, we observed that groups had different degrees of variation. The two falcata groups had more sequence variation than either caerulea group or hemicycla. Both caerulea and hemicycla groups contained ~45% of the amount of variation found in upland falcata, and <40% of that found in lowland falcata. A similar pattern of relative genetic diversity in lower magnitudes was observed in the comparisons of θ values between the five groups. Both caerulea groups and the hemicycla contained between 42-56% of the amount of variation found in either falcata groups (Table 4.3).

Linkage Disequilibria

Physical locations of some SSR loci were identified using the most recent *M. truncatula* genome sequence build (version 2), which is only about 2/3 complete (N. Young, pers. comm.). We have information on the location of 56 SSR loci and 185 locus pairs are known to reside on the same chromosome (Figures 4.1 and 4.2). The average distance between markers was 10.3Mb.

We found extensive LD among SSR markers that are on the same chromosome as well as those on different chromosomes. Over all genotypes, 73 locus pairs showed significant LD when all SSR markers were considered, but this number was reduced when the five groups were analyzed separately. Some of this reduction could be due to reduced power to detect LD when fewer genotypes were evaluated. Although upland falcata and southern caerulea had the same number of individuals, the number of SSR locus pairs in LD was much higher in upland falcata. The number of locus pairs in LD in hemicycla was equal to that of northern caerulea, in spite of the former group including fewer genotypes (Table 4.1). Although the distance of LD differed among the five main groups of wild diploid accessions, in general it spans around 10Mb. Disregarding the five groups, long-range LD extending as long as 20Mb was observed (Figure 4.4).

We tested 16 polymorphic sites for LD in the F5H sequence (Figure 4.3). We plotted r^2 (Hill and Robertson, 1968) against distance between polymorphic sites in bp. We found that LD among sites within F5H gene extended about 200bp, after which it dramatically decayed (Figure 4.3).

Association mapping

Based on genomewide association analysis with 89 SSR loci and 17 phenotypic traits, we only found one weak association (FDR Q-value = 0.053) with total biomass yield in 2008. Although this SSR marker did not match anywhere in the current sequence build of the *M. truncatula* genome, it had previously been mapped to linkage group 8 in alfalfa (Julier et al., 2003; Sledge et al., 2003). The marker did not show LD with any of the SSR loci known to be located on the chromosome 8. The polymorphic site 582 of *F5H* (highlighted green in Figure

4.3) was associated with total biomass yield of 2008. The significance level based on FDR Q-value was 0.01.

DISCUSSION

We observed a high number of SSR marker pairs in LD but the numbers in our study are lower than those reported in maize and barley (Remington et al., 2001; Liu et al, 2003; Stich et al., 2005; Malysheva-Otto et al., 2006). The lower number of SSR locus pairs in LD could partially be due to FDR calculations we used to correct possible false positives arising from multiple testing of thousands of locus pairs. It could also be due to the nature of the plant material. Previous studies selected landraces or inbred lines had resulted from human selection, which could create LD (Jannink and Walsh, 2002), whereas our germplasm contained all wild accessions.

The extent of LD based on SSR markers was much longer than within gene LD estimated directly from DNA sequence, and this phenomenon has been reported in other species as well (Remington et al., 2001; Stich et al., 2005). The difference in the extent of LD between different marker systems could be due to the fact that SSR markers have higher mutation rates than SNPs and could have evolved more recently, leading to new LD that extends over longer distance (Remington et al., 2001; Jannink and Walsh, 2002).

Data on one gene is not sufficient to make general conclusions about within-gene LD throughout the genome. However, the extent of LD in *F5H* in diploid alfalfa is similar to that previously reported in maize (Tenaillon et al., 2001) and much smaller than in *Arabidopsis* (Nordborg, 2000; Nordborg et al., 2002; Kim et al., 2007). Given that both alfalfa and maize are allogamous species and *Arabidopsis* is autogamous, this pattern of LD decay is expected. The sharp decline in extent of LD in F5H could be due to the gene segment analyzed. Since the

sequence was only about 500 bp, the relatively few available polymorphisms between two sites more than 200bp apart could have caused the absence of LD.

We identified two markers associated with biomass yield, but no associations with other traits. Given the paucity of markers we examined, this lack of association is not surprising. We could not infer the physical location of the SSR marker that is associated with total biomass yield in 2008, because it did not match with any of the known *M. truncatula* sequence. However it has been used in two different mapping studies (Julier et al., 2003; Sledge et al., 2005) and in both studies it was mapped to Linkage Group 8. The nature of this association needs to be further investigated.

The *F5H* gene was selected as one of several candidate genes for association of lignin and perhaps other cell wall constituents, but SNP variants in that gene were not associated with any composition component. One of the possible reasons is that F5H is involved in one of the later steps of lignin biosynthesis, involved in the formation of S-lignin, one of the components of the overall lignin molecule. Variation in this gene will have a larger effect on lignin composition than on lignin content. Down-regulation of F5H in transgenic plants has little effect on lignin content (Chen and Dixon, 2007). Thus, it is possible that we could not detect small natural variations statistically. The SNP we tested are from the first exon (Figure 4.3); an intron of considerable size lies between the two exons and we have demonstrated rapid decay in LD. Therefore, if a sequence polymorphism residing in the second exon causes the variation, the extensive decay in LD over that distance would have prevented us from detecting an association.

In summary, in this paper we attempt to estimate both genomewide SSR and within gene SNP variation to determine the extent of LD in diploid alfalfa. Our results suggest that using SSR markers for a genomewide association study may not be feasible. We also conclude that

there is a strong family structure effect on LD and selection of germplasm will be crucial to reduce the effect of family structure. Association testing of more candidate genes as well as genomewide SNP variants is underway to find useful marker-trait associations that can be used in breeding programs.

REFERENCES

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horto, K. Jakob, C. Lister, J. Molitor, C. Shindo, Tang C., et al. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1: e60.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Chen, F., and R.A. Dixon. 2007. Lignin modification improves fermentable sugar yields for biofuel production. Nature Biotechnology 25: 759-761.

Crews, T.E. and M.B. Peoples. 2004. Legume versus fertilizer sources of nitrogen: Ecological tradeoffs and human needs. Agriculture, Ecosystems, and Environment 102:279-297.

Delong, M.M., D.R. Swanberg, E.A. Oelke, C. Hanson, M. Onischak, M.R. Schmid, and B.C. Wiant. 1995. Sustainable biomass energy production and rural economic development using alfalfa as a feedstock. P. 1582–1591. *In* D.L. Klass (ed.) Second Biomass Conf. of the Americas: Energy, Environment, Agriculture, and Industry, Portland, OR. 21–24 Aug.

Evanno, G., S. Regnaut, J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611–2620.

Flint-Garcia, S.A., J.M. Thornsberry and E.S. Bucker. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54: 357–374.

Hagenblad, J. and M. Nordborg. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. Genetics 161:289–98.

Hardy, O. J. and X. Vekemans. 2002. SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620.

Hartl, D.L. and A.G. Clark. 1997. Principles of population genetics. 3rd edition. Sinauer Associates , Sunderland, MA.

Hill, W.G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226 - 231.

Hill, R.R., Jr., J.S. Shenk, and R.F Barnes. 1988. Breeding for yield and quality. p. 809–825. *In* A.A. Hanson et al. (ed.) Alfalfa and alfalfa improvement. ASA–CSSA–SSSA, Madison, WI.

Jannink, J. L., and B. Walsh. 2002. Association mapping in plant populations, pp. 59–68 in Quantitative Genetics, Genomics and Plant Breeding, edited by M. S. Kang. CAB International, NY.

Julier, B., S. Flajoulot, P. Barre, G. Cardinet, S. Santoni, T. Huguet, and C. Huyghe. 2003. Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago sativa*) using microsatellite and AFLP markers. BMC Plant Biol. 3:9.

Kaló, P., G. Endre, L. Zimányi, G. Csanádi, and G.B. Kiss. 2000. Construction of an improved linkage map of diploid alfalfa (Medicago sativa). Theor. Appl. Genet. 100:641–657

Kim, S., V. Plagnol, T.T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J.R. Ecker, D. Wiegel, and M. Nordborg. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39: 1151–1155.

Lamb, J.F.S., H.J.G. Jung, C.C. Sheaffer, and D.A. Samac. 2007. Alfalfa leaf protein and stem cell wall polysaccharide yields under hay and biomass management systems. Crop Sci. 47:1407–1415.

Lande, R. and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756.

Lander, E.S. and N.J. Schork. 1994. Genetic dissection of complex traits. Science 265: 2037–2048.

Liu, K.J., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128.

Mackay, I. and W. Powell. 2007. Methods for linkage disequilibrium mapping in crops. Trends in Plant Science Vol.12 No.2.

Malysheva-Otto, L.V., M.W. Ganal, M.S. Röder. 2006. Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genet. 7:6.

McCaslin, M. and D. Miller. 2007. The future of alfalfa as a biofuels feedstock. 37th California Alfalfa and Forage Symposium. December 17-19 2007, Monterey, CA.

Michaud, R., W.F. Lehman and M.D. Rumbaugh. 1988. World distribution and historical development. *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Neale, D.B., and O. Savolainen. 2004. Association genetics of complex traits in conifers. Trends Plant Sci. 9:325-330.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self fertilization. Genetics 154:923–29.

Nordborg, M., J.O. Borevitz, J. Bergelson, C.C. Berry, J. Chory et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 30:190–93.

Oraguzie, N.C., P.L. Wilcox, E.H.A. Rikkerink, and H.N. deSilva. 2007. Linkage disequilibrium in Oraguzie et al. (ed) Association mapping in plants pp 3-39. Sprigerleng, NY.

Patzek, T.W. 2004. Thermodynamics of the corn-ethanol biofuel cycle. Crit. Rev. Plant Sci. 23:519-567.

Pritchard, J.K., M. Stevens, P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Quiros, C.F., and G.R. Bauchan. 1988. The genus *Medicago* and the origin of the *Medicago sativa* complex, p. 93-124, *In* A. A. Hanson, et al., eds. Alfalfa and alfalfa improvement. ASA-CSSA-SSSA, Madison, WI.

Rafalski, A., and M. Morgante. 2004. Corn and humans: Recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20:103–111.

Ragauskas, A.J., C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, Jr., J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer, and T. Tschaplinski. 2006. The path forward for biofuels and biomaterials. Science 311:484-489.

Raymond, M., F. Rousset. 1995. GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. J. Heredity 86:248-249.

Reddy, M.S.S., F. Chen, G. Shadle, L. Jackson, H. Aljoe, and R.A. Dixon. 2005. Targeted down-regulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (*Medicago sativa* L.). Proc. Natl. Acad. Sci. USA 102:16573–16578.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98: 11479–11484.

Ritland, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. 67:175–185.

Rozas, J., J.C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas. 2003**.** DnaSP, DNA polymorphism analyses by the coalescent and other methods**.** Bioinformatics 19:2496-2497.

SAS Institute Inc., 2004. SAS® 9.1.2. 2004. Qualification tools user's guide. Cary, NC: SAS Institute Inc.

Sledge, M.K., I.M. Ray, and G. Jiang. 2005. An expressed sequence tag SSR map of tetraploid alfalfa (*Medicago sativa* L.). Theor. Appl. Genet. 111:980–992.

Stephens, M. and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics, 73: 1162-1169.

Stephens, M., N. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68: 978-989.

Stich, B., A.E. Melchinger, M. Frisch, H.P. Maurer, M. Heckenberger, and J.C. Reif. 2005. Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor. Appl. Genet. 111:723–730.

Storey, J.D. 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B 64: 479–498

Storey, J.D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100: 9440−9445.

Stuber, C.W., M. Polacco, and M.L. Senior. 1999. Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. Crop Sci. 39:1571–1583.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA 98:9161–9166.

U.S. DOE. 2006. Breaking the biological barrier to cellulosic ethanol: A joint research agenda, DOE/SC-0095, U.S. Department of Energy Office of Science and Office of Energy Efficiency and Renewable Energy (www.doegenomestolife.org/biofuels/)

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.

Yu, J.M., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Table 4.1. Number of SSR locus pairs showing linkage disequilibrium in five main populations of diploid alfalfa and over all genotypes based on a significance level of P = 0.01after control for the false discovery rate (FDR).

Groups	No. of genotypes	No. of locus pairs in LD	% of locus pairs in LD
Southern caerulea	99	37	0.9
Northern caerulea	69	21	0.5
Hemicycla	44	21	0.5
Lowland falcata	63	50	1.3
Upland falcata	99	47	1.2
Overall	374	73	1.8

Table 4.2. Summary of DNA sequence variation in the *F5H* gene in the five main populations of diploid alfalfa and across all genotypes.

Population	No. of genotypes	No. of poly-morphic sites	$\pi^{\dagger}$	θ
Southern caerulea	56	12	0.0054	0.0081
Northern caerulea	44	14	0.0061	0.0099
Hemicycla	23	9	0.0056	0.0073
Lowland falcata	37	24	0.0150	0.0177
Upland falcata	46	25	0.0126	0.0176
Overall	206	38	0.0126	0.0207

† π = the average pairwise difference between sequences (Tajima, 1983), and θ = Watterson's estimator of the sequence diversity (Watterson, 1975)

Table 4.3**.** Significant marker-trait associations after correction for multiple testing using the positive FDR method (*Q*-values).

Trait	Maker name	Marker type	Linkage group	Marker effect		FDR Q-value
				F	P	
Yield, 2008	*MTIC103*	SSR	8	2.5	0.00009	0.053
Yield, 2008	F5H site582	*SNP*	5	11.0	0.00003	0.01

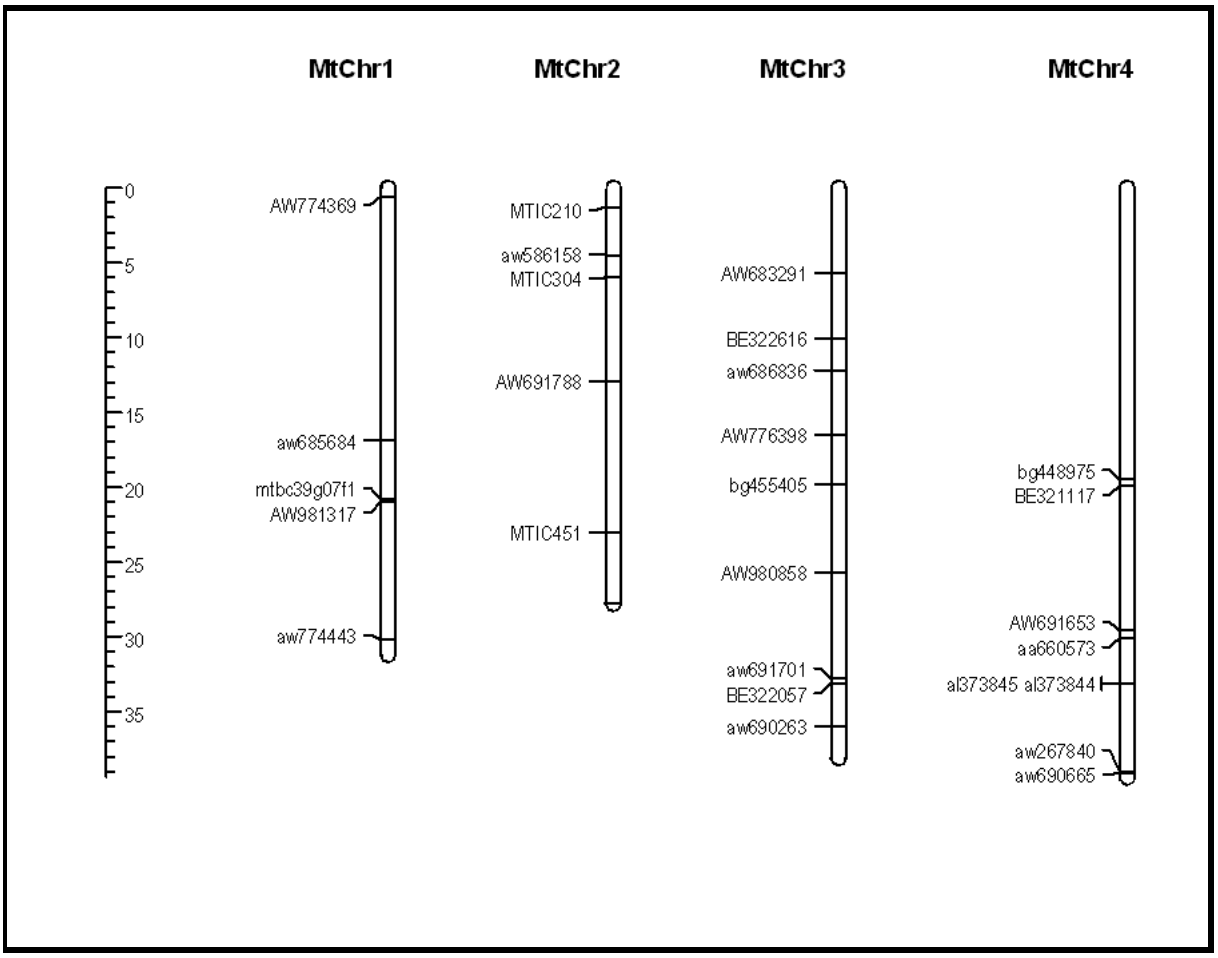

Figure 4.1. Physical location of SSR markers on *M. truncatula* chromosomes1-4. The ruler indicates their respective position from one end in the scale of Mb.

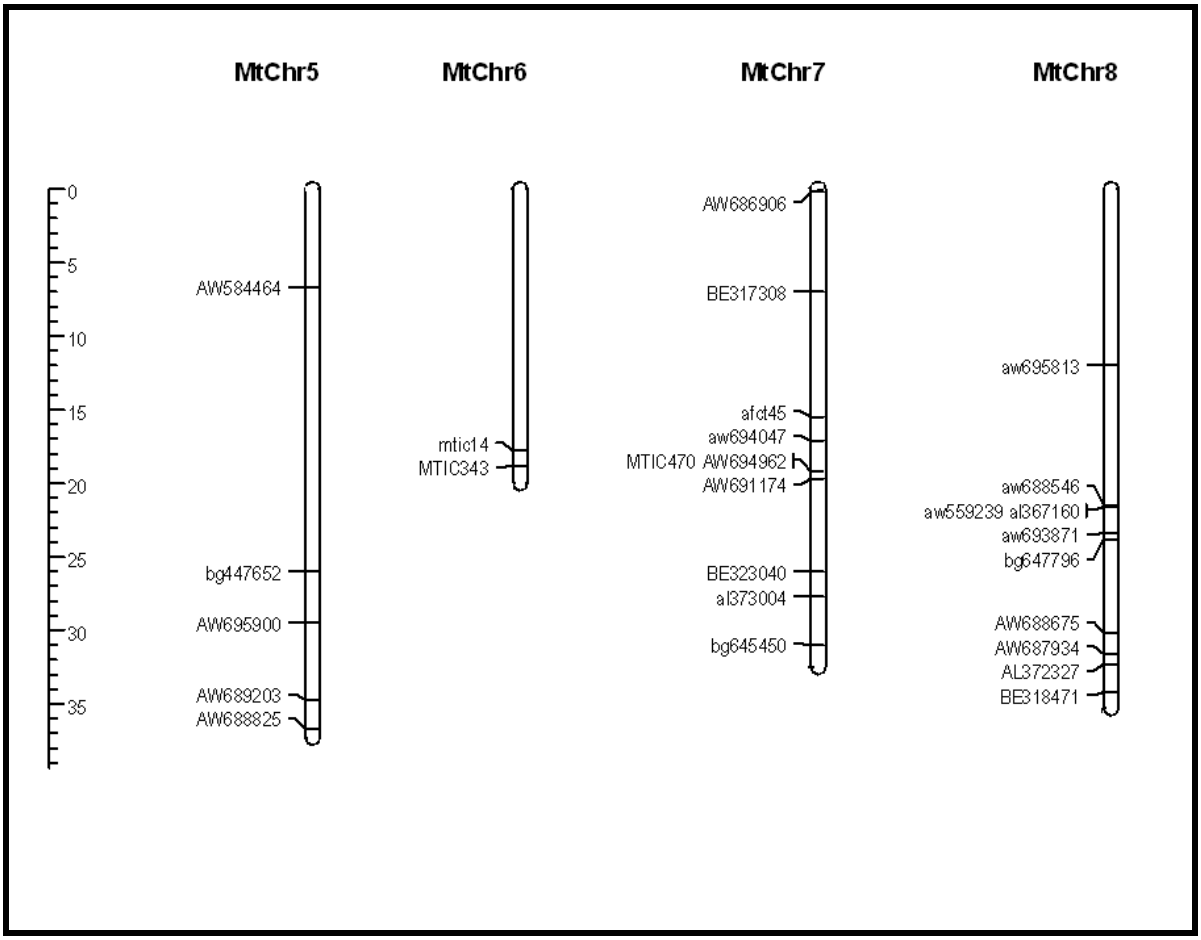

Figure 4.2. Physical location of SSR markers on *M. truncatula* chromosomes 5-8. The ruler indicates their respective position from one end in the scale of Mb.

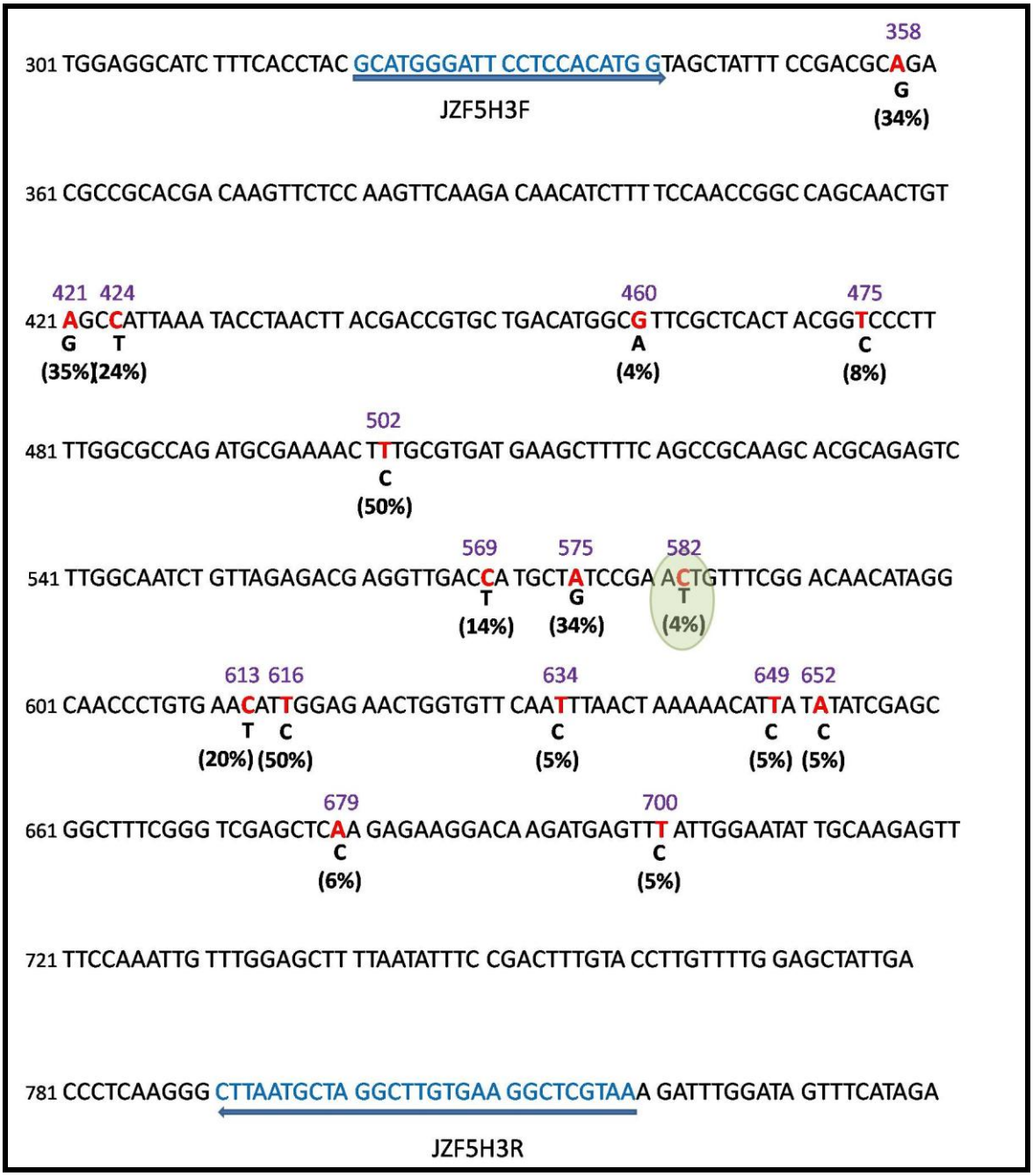

Figure 4.3. The consensus sequence of the fragment of F5H that was amplified and sequenced in 206 diploid alfalfa genotypes. The positions of SNPs among genotypes are indicated in red; the alternative nucleotide and its frequency at each location are indicated below the consensus sequence. Nucleotide 582 (circled) is associated with yield in 2008. SNPs with frequency less than 4% are not shown.

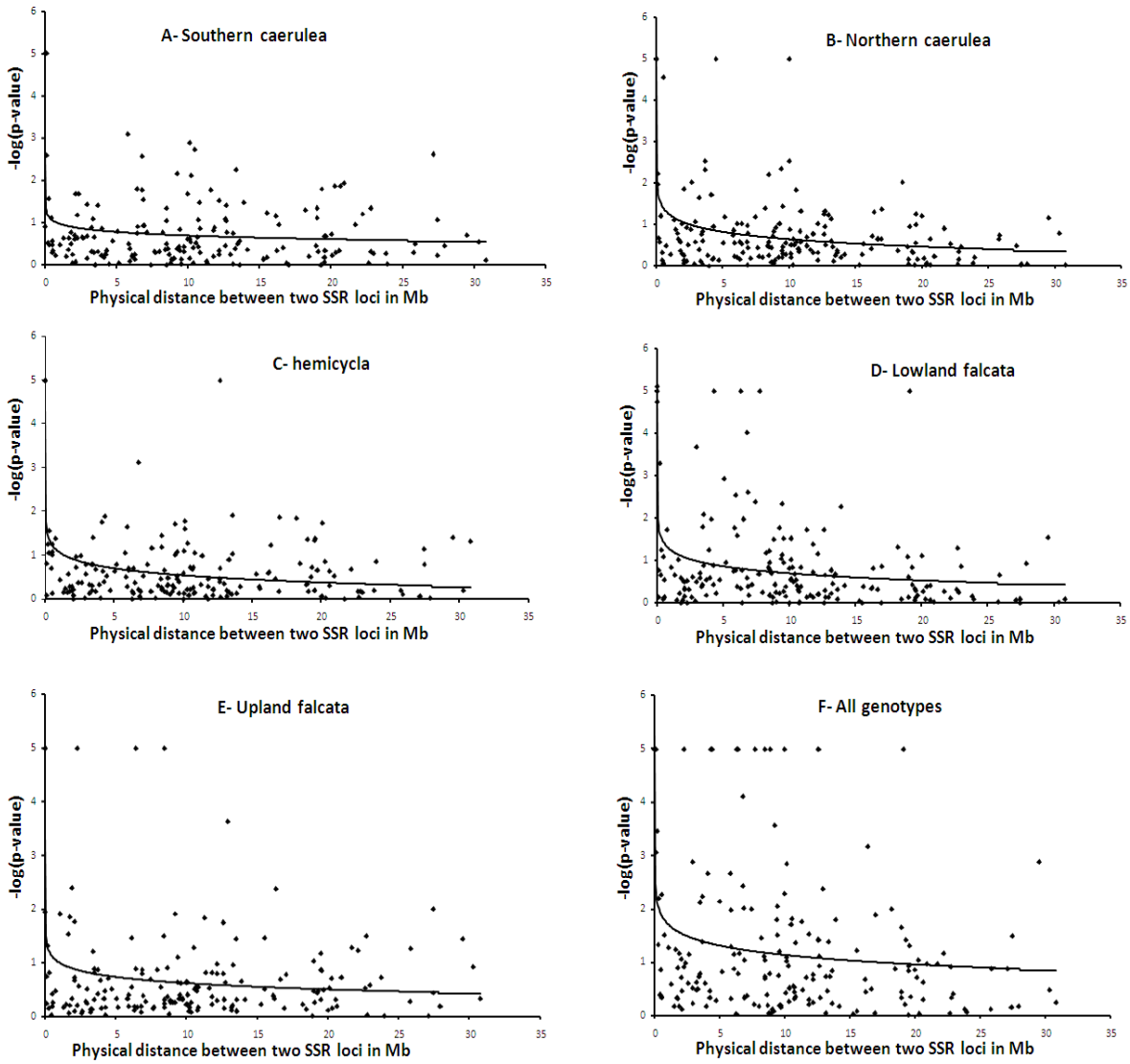

Figure 4.4. Plots of linkage disequilibrium (–log(p-value)) between SSR locus pairs on the same chromosome against their physical distance in Mb, based on the *M. truncatula* genome sequence, in five diploid alfalfa populations and over all genotypes.

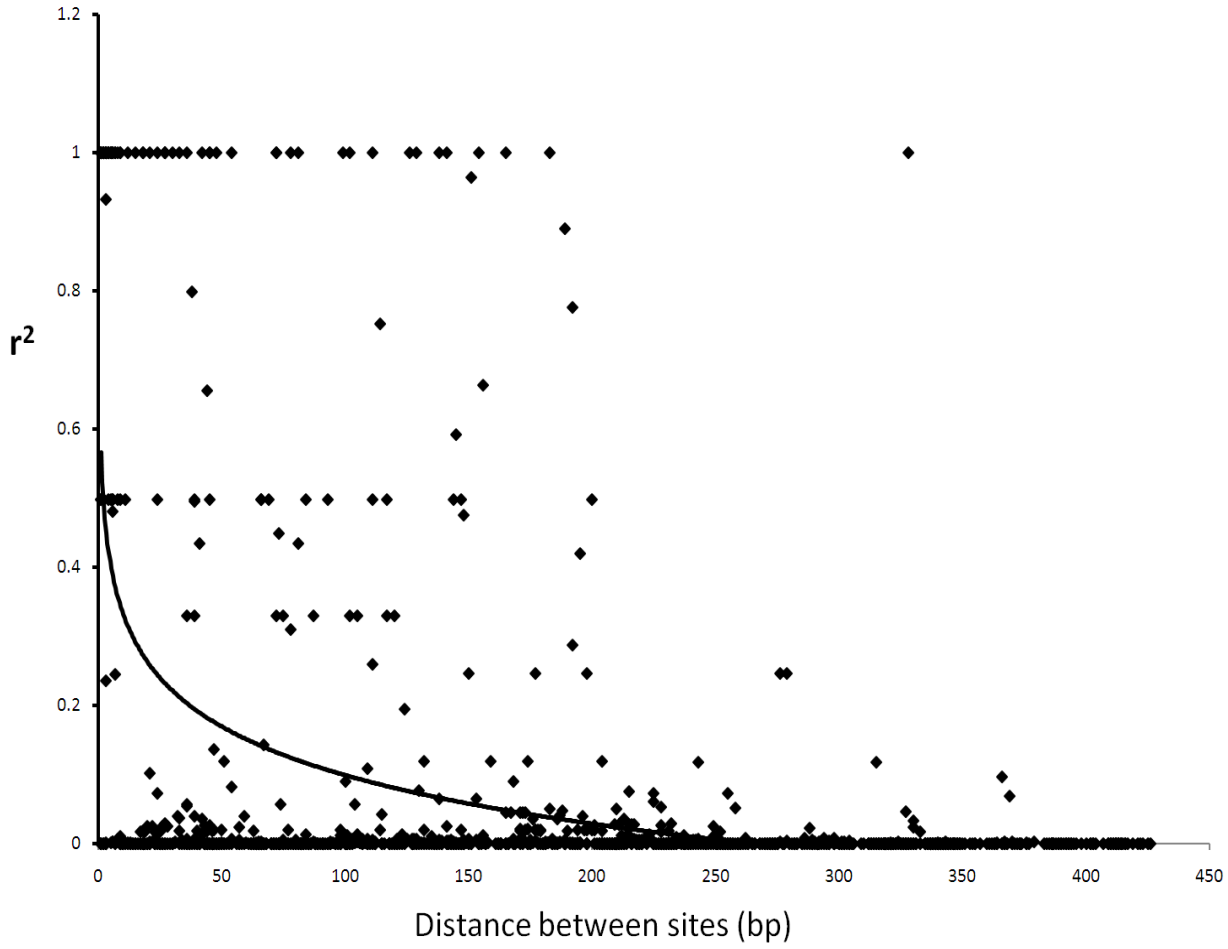

Figure 4.5. Plot of the squared correlation of allele frequencies (r^2) against the distance between polymorphic sites (bp) in the F5H gene across 206 wild diploid genotypes.

CHAPTER 5

CONCLUSIONS

In this study, I conducted the first large scale examination of the genetic and phenotypic diversity of diploid alfalfa germplasm. Despite the simpler inheritance of diploids and the wealth of genetic diversity present among the germplasm, virtually no investigations have previously been conducted to quantify the variation within this germplasm.

We selected 374 genotypes from 120 wild diploid accessions throughout the natural distribution range of *Medicago sativa* subspecies *falcata*, caerulea, and *hemicycla* and used SSR markers to examine the genetic structure of diploid taxa. The SSR markers clearly separated the subspecies falcata and caerulea and the hybrid nature of hemicycla was confirmed. We also observed hierarchical population structure suggesting further separation of falcata and caerulea into two ecotypes. The two caerulea groups largely correspond to Northern and Southern regions of Eurasia. The separation of falcata accessions into two groups was based on ecogeograpy. The first group, which we termed the *lowland* ecotype, consisted of accessions from lowlands and locations closer to river basins or the sea coast. We termed the second group the *upland* ecotype, as accessions were collected from dry slopes of mountain ranges and/or high elevations. At this point, this distinction is tentative since many accessions had no passport information on the exact place of origin.

The evaluation of genetic diversity also led us to conclude that most of the genetic variation is within populations, even though we used few individuals within accessions. We have also found that an immense amount of genetic variation exists in diploid gene pool that could be

used to broaden the cultivated alfalfa genepool. The identification of different ecotypes could help breeders to more effectively use gene pools.

We also evaluated cell wall constituents and agronomic traits of selected germplasm and observed wide variation among individuals and accessions for all the traits evaluated. Since lignin is limiting factor for effective saccarification of cell wall sugars (Chen and Dixon, 2007), and high lignin results in lower forage nutritive value and lower biofeedstock quality (Buxton and Russell, 1987; Albrecht et al., 1988; Reddy et al 2005), low lignin is a desirable trait for cellulosic ethanol production or for animal nutrition. We found that the lignin level of the unimproved diploid alfalfa stems ranged from 61 to 123 g/kg with the mean of 100 g/kg. The extent of variation of cell wall components in our study of diploids was 2-4 fold larger compared to those reported in tetraploid germplasm core collection (Jung et al., 1997). This implies that diploid germplasm could yield genotypes with unique combinations of cell wall composition – lower lignin and higher cellulose content – that could be incorporated into tetraploid breeding programs via unreduced gametes. Our results of individual genotypes could also be used for identifying the best parental combinations (e.g., extreme phenotypes) for mapping the genetic control of forage quality, biofuel potential, and yield in diploid accessions. Mapping directly at the diploid level will eliminate complications arising from tetrasomic inheritance.

We were interested in testing if the five groups found based on genomic structure could also be observed based on phenotype. Despite the differences observed in trait means of all five groups, the phenotypic data did not follow a pattern of separation similar to the molecular marker results. The lack of a relationship between molecular marker based clustering and phenotypic clustering implies that desirable alleles for improvement of yield and cell wall traits (either in the

context of forage quality or biofuel potential) are dispersed across all subspecies, potentially providing a diversity of alleles for the traits.

We observed a high number of SSR marker pairs in LD with a distance that extends as long as 20 Mb. We also found that within gene LD extends up to 200bp, but dramatically decays afterward 200 bp. The extent of LD based on SSR markers was much longer than within gene LD estimated directly from DNA sequence. The difference in the extent of LD between different marker systems could be due to the fact that SSR markers have higher mutation rates than SNPs and could have evolved more recently, leading to new LD that extends over longer distance (Remington et al., 2001; Jannink and Walsh, 2002). Our results suggest using SSR markers for a genomewide association study may not be feasible. We also conclude that there is a strong family structure effect on LD and selection of germplasm will be crucial to reduce the effect of family structure. Association mapping of more candidate genes is needed to test the applicability of a candidate gene approach. A test of genomewide SNP markers, once developed, could help to determine if genomewide associations could be detected.

REFERENCES

Albrecht, K. A., W. F. Wedin, and D. R. Buxton. 1987. Cell-wall composition and digestibility of alfalfa stems and leaves. Crop Sci. 27:735–741.

Buxton, D. R. and J. R. Russell. 1988. Lignin constituents and cell-wall digestibility of grass and legume stems. Crop Sci. 28:553–558.

Chen, F. and R.A. Dixon. 2007. Lignin modification improves fermentable sugar yields for biofuel production. Nature Biotechnology 25:759-761.

Jannink, J. L. and B. Walsh. 2002. Association mapping in plant populations, pp. 59–68 in Quantitative Genetics, Genomics and Plant Breeding, edited by M. S. Kang. CAB International, NY.

Jung, H.G., C.C. Sheaffer, D.K. Barnes, and J.L. Halgerson. 1997. Forage quality variation in the U.S. alfalfa core collection. Crop Sci. 37:1361-1366.

Reddy, M.S.S., F. Chen, G. Shadle, L. Jackson, H. Aljoe, and R.A. Dixon. 2005. Targeted down-regulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (*Medicago sativa* L.). Proc. Natl. Acad. Sci. USA 102:16573–16578.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484.