

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Phenotypic and Genetic Characterization of Non-Dormant Alfalfa Germplasm Across Diverse Environments using DArTag Markers

Permalink

<https://escholarship.org/uc/item/4z0517hm>

ISBN

9798189237690

Author

Nonkum, Kreingkrai

Publication Date

2026-06-25

Supplemental Material

<https://escholarship.org/uc/item/4z0517hm#supplemental>

Peer reviewed|Thesis/dissertation

Phenotypic and Genetic Characterization of Non-Dormant Alfalfa Germplasm Across
Diverse Environments using DArTag Markers

By

KREINGKRAI NONKUM
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Horticulture and Agronomy

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

E. Charles Brummer, Chair

Daniel Putnam

Allen Van Deynze

Committee in Charge

2026

Abstract

This dissertation investigates the genetic diversity of non-dormant alfalfa (*Medicago sativa* L.) using a mid-diversity 3K DArTag platform. Maintaining allelic diversity within germplasm collections is a crucial prerequisite for effective alfalfa breeding. Chapter 1 reviews the significance of non-dormant alfalfa production in Mediterranean climates and the utility of DArTag technology for genomic and genetic diversity studies. Chapter 2 uses the DArTag platform to assess the genetic diversity of 376 bulked GRIN accessions from seven geographic origins using 2,769 SNP loci harboring microhaplotypes. Microhaplotypes reduced monomorphic loci from 33% to 4% relative to target SNPs, with high genotyping reproducibility across replicate pools (mean $r^2 = 0.92$). STRUCTURE analysis identified five clusters corresponding to geographic regions, with AMOVA confirming significant among-population differentiation. Chapter 3 evaluates DArTag markers for genome-wide association studies (GWAS) and genomic prediction within multi-environmental trials across three locations. A stagewise with P-splines spatial model yielded adjusted phenotypic means for dry weight, plant height, regrowth, and persistence. GWAS identified 69 marker-trait associations shared across all four traits. Genomic prediction (GBLUP) yielded predictive abilities of 0.17–0.75 for dry weight and 0.54–0.84 for plant height across harvest environments. Chapter 4 applies DArTag markers to cultivar identification across 84 seed lots of “CUF 101” and check cultivars using 1,588 SNPs. PCA distinguished CUF 101 from check cultivars, while DAPC resolved three genetic clusters within CUF 101 seed lots; one cluster of certified lots showed substantial deviation from the core genetic profile, suggesting contamination or admixture linked to seed company source and certification year. Overall, this dissertation demonstrates the versatility of the DArTag platform for genetic diversity analysis, GWAS, genomic prediction, and cultivar identification, providing a robust framework for modern alfalfa breeding programs.

Contents

Chapter 1: Literature Review.....	1
Chapter 2 : Analysis of Non-Dormant Alfalfa (<i>Medicago sativa</i>) Genetic Diversity and Population Structure using Genome Wide Microhaplotypes Markers.....	14
1. Introduction.....	15
2. Materials and Methods.....	18
3. Results.....	22
4. Discussion.....	41
5. Conclusion	49
6. Acknowledgement	50
7. References.....	50
Chapter 3 : Identifying the Genetic Basis of Four Agronomic Traits in a Non-Dormant Alfalfa Germplasm Collection	58
1. Introduction.....	59
2. Materials and Methods.....	63
3. Results.....	72
4. Discussion.....	89
5. Conclusion	101
6. Acknowledgement	102
7. References.....	102

Chapter 4 : Genetic Variability Among Seed Lots of the Alfalfa Cultivar ‘CUF 101’	128
1. Introduction.....	129
2. Materials and Methods.....	131
3. Results.....	134
4. Discussion.....	140
5. Conclusion	142
6. Acknowledgement	142
7. References.....	142
Concluding Remarks.....	149

Chapter 1

Literature Review

1.1 Alfalfa

Alfalfa (*Medicago sativa* L.) is a perennial legume known for its high nutritional value and central role in livestock production. It serves as a key feed source for dairy and beef cattle as well as other livestock species, due to its high crude protein content, excellent digestibility, and palatability (Summers and Putnam, 2008). Native to regions spanning Siberia, the Arabian Peninsula, and North Africa (Michaud et al., 1988), alfalfa spread globally through the expansion of civilizations and the Silk Road trade routes (Wang & Şakiroğlu, 2021), eventually establishing itself across a wide range of agroecological zones. Owing to its broad adaptability, alfalfa is grown extensively in both temperate and dry tropical regions, including intermountain to low desert areas in California (Summers and Putnam, 2008). This broad adaptability, combined with its perennial growth habit and nitrogen-fixing capacity, underlies alfalfa's status as one of the most widely cultivated forage crops in the world.

North America, particularly the United States, remains the leader in alfalfa production and export. In 2025 the total value of alfalfa hay in the United States was estimated to exceed \$7 billion (NASS, 2025). However, the value of alfalfa hay does not capture the alfalfa's broader economic impact, as the crop also contributes indirectly to the value of livestock products such as dairy products from the animals fed on alfalfa-based feed (Summers and Putnam, 2008).

Beyond its economic significance, Alfalfa offers important agronomic and ecological benefits. As a leguminous species capable of biological nitrogen fixation, alfalfa enriches soil fertility and reduces the need for synthetic nitrogen fertilizers (Vance et al., 1988). Its perennial growth habit provides year-round ground cover, reducing soil erosion, enhancing water retention, and promoting biodiversity in agricultural landscapes. Furthermore, alfalfa contributes to long-term soil health by increasing soil organic carbon and enhancing carbon sequestration (Fernandez et al., 2019).

1.2 Alfalfa Genetics and Genomics

Alfalfa (*Medicago sativa* L.) is an autotetraploid species within the *Medicago* genus with a chromosome number of $2n = 4x = 32$, while its diploid relatives possess $2n = 2x = 16$ (McCoy and Bingham, 1988; Quiros and Bauchan, 1988). Most commercial cultivars are derived from *M. sativa* subsp. *sativa*, which typically exhibits purple flowers, whereas *M. falcata*, characterized by yellow flowers, contributes alleles associated with winter hardiness and cold tolerance (Teuber and Brick, 1988; Lesins & Lesins, 1979). The introgression of *M. falcata* germplasm into *M. sativa* breeding populations has enhanced adaptability to harsh environments and extended the range of cultivation into colder regions. This broad genetic base, however, also reflects a fundamental reproductive characteristic that profoundly shapes alfalfa breeding strategy.

Alfalfa is an outcrossing species due to strong self-incompatibility, which prevents self-fertilization and maintains high levels of genetic diversity within populations. This outcrossing nature, while beneficial for diversity, means that repeated self-pollination results in severe inbreeding depression, leading to significant reductions in vigor, fertility, seed set, and overall yield (Busbice et al., 1972; Bingham et al., 1994; Li & Brummer, 2012). The negative effects of inbreeding are particularly pronounced in tetraploid alfalfa, where deleterious allelic interactions

decay more rapidly than in diploid species, resulting in a more rapid loss of fitness (Li & Brummer, 2012). These reproductive constraints have directly shaped the breeding methods used in commercial alfalfa improvement.

Current alfalfa breeding programs rely on recurrent phenotypic selection, a cyclical process designed to accumulate favorable alleles by inter-mating superior genotypes and forming synthetic populations (Li & Brummer, 2012). This approach circumvents the inbreeding problem by maintaining population-level outcrossing and has proven effective for improving agronomic traits such as biomass yield and forage quality (Hill et al., 1988), abiotic stress tolerance (Elgin et al., 1988), and pest resistance (Sorensen et al., 1988). The rate of yield improvement in alfalfa has been relatively slow compared with diploid grain crops such as maize, owing not only to the autotetraploid genome and complex polyploid inheritance which limit the precision of selection and the prediction of offspring performance but also to constraints nature of a perennial forage: the partitioning of assimilates between harvestable shoots and the root reserves needed for persistence and winter survival, the absence of a harvest-index lever analogous to grain, and the continual compromise between yield and forage quality across maturity and cutting schedules (Bingham et al., 1994; Orloff and Putnam, 2007; Volenec et al., 1996).

To overcome these constraints, researchers have proposed exploiting heterosis (hybrid vigor) by developing and crossing genetically distinct breeding populations (Brummer, 1999). The identification of heterotic groups in alfalfa is a promising avenue for yield improvement, yet it remains an active area of investigation (Rumbaugh et al., 1988; Riday and Brummer, 2005; Li et al., 2009). A better understanding of heterotic group structure, combined with advances in molecular breeding and genomic prediction, is expected to accelerate the development of high-yielding cultivars, particularly because the complexity of alfalfa's genome is precisely what makes genome-informed approaches so valuable.

1.3 Fall dormancy

Fall dormancy (FD), also referred to as autumn dormancy, is a physiological response which reduces or stops continued winter growth of alfalfa in response to environmental conditions (McKenzie et al., 1988). This adaptive response is primarily triggered by decreasing photoperiod and temperature during the autumn months (Castonguay et al., 2006). In alfalfa, fall dormancy is an important trait used to classify cultivars based on their regrowth potential after a fall cutting. The standard method for assessing fall dormancy involves measuring plant height after regrowth in late summer or early fall and converting these values into standardized dormancy ratings relative to check cultivars (Teuber et al., 1998).

Dormancy ratings in alfalfa are expressed on a numerical scale from 1 to 11. A score of 1 represents very dormant types with limited fall regrowth, and 11 indicates non-dormant cultivars exhibiting vigorous fall and winter growth (National Alfalfa and Forage Alliance, 2025). Based on this system, cultivars are classified as follows: 1-2 (very dormant), 3-4 (dormant), 5 (moderately dormant), 6-7 (semi-dormant), 8-9 (non-dormant), and 10-11 (very non-dormant). Dormancy level selection is region-specific; non-dormant cultivars are best suited for regions with mild winters where extended growth maximizes yield, whereas dormant cultivars are preferred in high-latitude or high-elevation regions with harsh winters, providing superior winter survival and forage quality, but typically with lower total yield (Summers and Putnam, 2008).

The improvement of fall growth while maintaining winter hardiness remains a major breeding objective. Historically, these two traits were considered negatively correlated, plants with greater fall growth were believed to be more susceptible to winter injury. This perceived trade-off has posed a long-standing challenge for breeders attempting to develop cultivars that sustain high

productivity while enduring cold winter conditions. However, recent studies indicate that fall growth and winter hardiness may not be perfectly genetically correlated. The current research suggests that it is possible to enhance fall growth potential without compromising winter survival, thereby breaking the trade-off between fall regrowth and winter hardiness (Brummer et al., 2000).

1.4 Germplasm collection

Genetic diversity represents a fundamental resource for plant breeding, populations with high intrinsic variation provide greater opportunities to develop improved cultivars (Govindaraj et al., 2015). Access to diverse plant germplasm enables breeders to deploy two complementary strategies which are introgression and incorporation (Simmonds, 1993). Introgression involves backcrossing of exotic or wild germplasm into elite populations to introduce discrete desirable traits, particularly those of high heritability controlled by one or a few genes. In contrast, incorporation is a long-term strategy in which exotic germplasm is first adapted to the target environment and then incorporated into the breeding base, allowing accumulation of small-effect alleles for quantitative trait improvement (Simmonds, 1993).

The establishment and maintenance of germplasm collections are essential for preserving allele diversity that may otherwise be lost during intensive selection in breeding programs. Genetic diversity tends to decline with repeated selection of elite cultivars, a trend that reduces the potential for further genetic gain and raises the risk of genetic vulnerability (Sanchez et al., 2023). In the case of alfalfa, this issue is particularly acute among non-dormant cultivars, which often exhibit lower genetic diversity compared to dormant types (Grandón et al., 2013). Reductions in heterosis among accessions (Bhandari et al., 2011) and smaller observed diversity in elite cultivars relative to wild populations (Müller et al., 2006) further underline the need for broad alfalfa germplasm bases. Effectively

harnessing this diversity requires robust tools for characterizing genetic variation across large collections, a need that has driven major advances in molecular marker technology.

1.5 Molecular markers for genetic diversity study in alfalfa

Molecular markers have become indispensable tools for exploring the plant genome diversity and facilitating marker-assisted selection (MAS) in crop improvement programs. In alfalfa early genetic studies predominantly relied on Simple Sequence Repeat (SSR) because of their codominant inheritance, high reproducibility, and multi-allelic nature (Grandón et al., 2013, Bhandari et al., 2011, Annicchiarico et al., 2015). Despite their advantages, genotyping with SSR markers is labor-intensive and typically yields a limited number of loci, often fewer than 100, which constrains genome-wide analysis and breeding applications.

The advent of next generation sequencing technology (NGS) has revolutionized genotyping approaches by enabling high throughput detection of thousands of single nucleotide polymorphisms (SNPs) across the genome. Genotyping platforms such as genotype by sequencing (GBS) (Elshire et al., 2011) have significantly enhanced markers density and genome coverage allowing more detailed assessments of population structure, linkage disequilibrium, and genetic diversity. For example, Li et al. (2014) developed an Illumina Infinium array containing 9,277 SNPs derived from the transcriptome sequencing, providing one of the first large-scale SNP resources for the alfalfa genome.

With the release of a high-quality reference genome for autotetraploid alfalfa (Chen et al., 2020; Shen et al., 2020), new opportunities have emerged for developing targeted species-specific genotyping platforms. One notable advancement is the DArTag (Diversity Arrays Technology) platform, developed through the Breeding Insight initiative (Zhao et al., 2023). This targeted

amplicon-based method improves upon traditional GBS by minimizing missing data and ensuring that all samples are sequenced at the same genomic loci. The DArTag platform currently targets approximately 3,000 SNPs distributed throughout the alfalfa genome, with each target SNP flanked by a sequencing region of 54–81 base pairs and a sequencing depth of 100×. Each target region produces a short sequence surrounding the SNP known as a microhaplotype, which contains multiple variable bases. The multi-allelic nature of microhaplotypes provides richer genetic information per locus than single SNPs, making them particularly advantageous for autotetraploid species like alfalfa. Their inclusion enhances the detection of allelic diversity, increases the intensity of information from markers, and improves the precision of downstream analyses such as population structure, phylogenetic clustering, and genomic prediction (Zhao et al., 2023; Medina et al., 2025).

Overall, the integration of comprehensive germplasm resources, high-density molecular markers, and advanced sequencing platforms has greatly improved the capacity for genetic diversity analysis and molecular breeding in alfalfa. As genomic resources continue to expand, these technologies will play a crucial role in accelerating cultivar development, enhancing selection efficiency, and preserving genetic variability within breeding programs.

References

- Annicchiarico, P., Nazzicari, N., Ananta, A., Carelli, M., Wei, Y., & Brummer, E. C. (2016). Assessment of cultivar distinctness in alfalfa: A comparison of genotyping-by-sequencing, simple-sequence repeat marker, and morphophysiological observations. *The Plant Genome*, 9(2), plantgenome2015-10.
<https://doi.org/10.3835/plantgenome2015.10.0105>
- Bhandari, H.S., Lawati, A.H.A. and Ray, I.M. (2011), Amplified Fragment Length Polymorphism Marker Diversity among High Yielding Perennial Alfalfa (*Medicago sativa* L.) Core Collection Accessions. *Crop Sci.*, 51: 1110-1121.
<https://doi.org/10.2135/cropsci2010.08.0492>
- Bingham, E.T., Groose, R.W., Woodfield, D.R. and Kidwell, K.K. (1994), Complementary Gene Interactions in Alfalfa are Greater in Autotetraploids than Diploids. *Crop Science*, 34: 823-829 cropsci1994.0011183X003400040001x.
<https://doi.org/10.2135/cropsci1994.0011183X003400040001x>
- Brummer, E.C. (1999), Capturing Heterosis in Forage Crop Cultivar Development. *Crop Science*, 39: 943-954 cropsci1999.0011183X003900040001x.
<https://doi.org/10.2135/cropsci1999.0011183X003900040001x>
- Brummer, E. C., Shah, M. M., & Luth, D. (2000). Reexamining the relationship between fall dormancy and winter hardiness in alfalfa. *Crop Science*, 40(4), 971-977.
<https://doi.org/10.2135/cropsci2000.404971x>

- Busbice, T.H., Hill, R.R., Jr. and Carnahan, H.L. (1972). Genetics and Breeding Procedures. In *Alfalfa Science and Technology*, 283-318. C.H. Hanson (Ed.).
<https://doi.org/10.2134/agronmonogr15.c13>
- Castonguay, Y., Laberge, S., Brummer, E. C., & Volenec, J. J. (2006). Alfalfa winter hardiness: a research retrospective and integrated perspective. *Advances in agronomy*, 90, 203-265.
- Elgin, J.H., Jr., Welty, R.E. and Gilchrist, D.B. (1988). Breeding for Disease and Nematode Resistance. In *Alfalfa and Alfalfa Improvement*, 29, 827 - 858. (eds A.A. Hanson, D.K. Barnes and R.R. Hill). <https://doi.org/10.2134/agronmonogr29.c27>
- Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., ... & Qiu, Q. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.*, 11(1), 2494.
<https://doi.org/10.1038/s41467-020-16338-x>
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5), e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Fernandez, A. L., Sheaffer, C. C., Tautges, N. E., Putnam, D. H., & Hunter, M. C. (2019). Alfalfa, wildlife, and the environment. *National Alfalfa and Forage Alliance*.
- Hill, R.R., Jr., Shenk, J.S. and Barnes, R.F. (1988). Breeding for Yield and Quality. In *Alfalfa and Alfalfa Improvement* (eds A.A. Hanson, D.K. Barnes and R.R. Hill).
<https://doi.org/10.2134/agronmonogr29.c26>

- Govindaraj, M., Vetriventhan, M. & Srinivasan, M. Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet Res Int* 1–14 (2015) <https://doi.org/10.1155/2015/431487>.
- Lesins K, Lesins I (1979) Genus *Medicago* (Leguminosae): A taxogenetic study. Kluwer, Dordrecht
- Grandón, N.G., Alarcón, Y., Moreno, M.V., Arolfo, V., Orodizzi, A., Basigalup, D.H., Gieco, J.O. and Bruno, C.I., 2013. Genetic diversity among alfalfa genotypes (*Medicago sativa* L.) of non-dormant cultivars using SSR markers and agronomic traits. *Revista de la Facultad de Ciencias Agrarias. Universidad Nacional de Cuyo*. 45(2) 45-59.
- Li, X., Wei, Y., Nettleton, D., & Brummer, E. C. (2009). Comparative gene expression profiles between heterotic and non-heterotic hybrids of tetraploid *Medicago sativa*. *BMC Plant Biology*, 9(1), 107.
- Li, X., & Brummer, E. C. (2012). Applied Genetics and Genomics in Alfalfa Breeding. *Agronomy*, 2(1), 40-61. <https://doi.org/10.3390/agronomy2010040>
- Li, X., Y. Han, Y. Wei, A. Acharya, A. D. Farmer, J. Ho, M. J. Monteros, and E. C. Brummer, (2014) Development of an alfalfa snp array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS One* 9: e84329.
- McCoy, T.J. and Bingham, E.T. (1988). Cytology and Cytogenetics of Alfalfa. In *Alfalfa and alfalfa improvement*, 29, 737 - 776. (eds A.A. Hanson, D.K. Barnes and R.R. Hill). <https://doi.org/10.2134/agronmonogr29.c24>

- McKenzie, J.S., Paquin, R. and Duke, S.H. (1988). Cold and Heat Tolerance. In *Alfalfa and Alfalfa Improvement*, 29, 737 - 776. (eds A.A. Hanson, D.K. Barnes and R.R. Hill).
<https://doi.org/10.2134/agronmonogr29.c8>
- Michaud, R., W. Lehman, and M. Rumbaugh, 1988 World distribution and historical development. *Alfalfa and alfalfa improvement* pp. 25-91. (eds A.A. Hanson, D.K. Barnes and R.R. Hill).
- Medina, C. A., Zhao, D., Lin, M., Sapkota, M., Sandercock, A. M., Beil, C. T., ... & Xu, Z. (2025). Pre-breeding in alfalfa germplasm develops highly differentiated populations, as revealed by genome-wide microhaplotype markers. *Scientific reports*, 15(1), 1253.
- Muller, M.-H., Poncet, C., Prosperi, J.M., Santoni, S. and Ronfort, J. (2006), Domestication history in the *Medicago sativa* species complex: inferences from nuclear sequence polymorphism. *Molecular Ecology*, 15: 1589-1602. <https://doi.org/10.1111/j.1365-294X.2006.02851.x>
- NASS, 2024 Nass statistics by subject. <http://tinyurl.com/nassalfalfa>, (Accessed: 09/12/2025).
- National Alfalfa and Forage Alliance (2025) *Alfalfa Variety Ratings Winter Survival Fall Dormancy & Pest Resistance Ratings for Alfalfa Varieties 2025 Edition*. Hay and Forage Grower. https://alfalfa.org/pdf/2025_Alalfa_Variety_Leaflet-high-res.pdf
- Orloff, S.B., Putnam, D.H. (2007). Harvest strategies for alfalfa. In: Summers, C.G., Putnam, D.H. (eds.), *Irrigated Alfalfa Management for Mediterranean and Desert Zones*, Ch. 13. UC ANR Publication 8299, Oakland, CA.
- Putnam, D.H., Orloff, S.B. and Teuber, L.R. (2007) *Irrigated Alfalfa Management: Choosing an Alfalfa Cultivar*. Ch. 5. ANR Pub., Oakland, 8291.

- Quiros, C. F., & Bauchan, G. R. (1988). The genus *Medicago* and the origin of the *Medicago sativa* comp. *Alfalfa and alfalfa improvement*, 29, 93-124. (eds A.A. Hanson, D.K. Barnes and R.R. Hill).
- Riday, H. and Brummer, E.C. (2005), Heterosis in a Broad Range of Alfalfa Germplasm. *Crop Science*, 45: 8-17 cropsoci2005.0008. <https://doi.org/10.2135/cropsoci2005.0008a>
- Rumbaugh, M.D., Caddel, J.L. and Rowe, D.E. (1988). Breeding and Quantitative Genetics. In *Alfalfa and Alfalfa Improvement*, 29, 777-808. (eds A.A. Hanson, D.K. Barnes and R.R. Hill). <https://doi.org/10.2134/agronmonogr29.c25>
- Sanchez, D., Sadoun, S.B., Mary-Huard, T., Allier, A., Moreau, L., & Charcosset, A. (2023). Improving the use of plant genetic resources to sustain breeding programs' efficiency, *Proc. Natl. Acad. Sci. U.S.A.* 120 (14) e2205780119, <https://doi.org/10.1073/pnas.2205780119>
- Shen, C., Du, H., Chen, Z., Lu, H., Zhu, F., Chen, H., ... & Wang, T. (2020). The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Molecular Plant*, 13(9), 1250-1261.
- Simmonds, N. W. (1993). Introgression and incorporation. Strategies for the use of crop genetic resources. *Biol. Rev. Camb. Philos. Soc.* **68**, 539–562
- Sorensen, E.L., Byers, R.A. and Horber, E.K. (1988). Breeding for Insect Resistance. In *Alfalfa and Alfalfa Improvement*, 29, 859-902. (eds A.A. Hanson, D.K. Barnes and R.R. Hill). <https://doi.org/10.2134/agronmonogr29.c28>
- Summers, C. G., & Putnam, D. H. (Eds.). (2008). *Irrigated alfalfa management for Mediterranean and desert zones* (Vol. 3512). UCANR Publications.

- Teuber, L.R.; Skrdla, W.H.; Beard, D.F.; Hunt, O.J.; Murphy, R.P.; Bingham, E.T.; Barnes, D.K. *Alfalfa Germplasm in the United States: Genetic Vulnerability, Use, Improvement, and Maintenance*; 1977; USDA Tech. Bull. No. 1571; U.S. Gov. Print. Office: Washington, DC, USA.
- Teuber, L.R. and Brick, M.A. (1988). Morphology and Anatomy. In *Alfalfa and Alfalfa Improvement*, 29, 125-162. (eds A.A. Hanson, D.K. Barnes and R.R. Hill).
<https://doi.org/10.2134/agronmonogr29.c4>
- Teuber, L. R., Taggard, K. L., Gibbs, L. K., McCaslin, M. H., Peterson, M. A., & Barnes, D. K. (1998). Fall dormancy. *Standard tests to characterize alfalfa cultivars*, A-1.
- Vance, C.P., Heichel, G.H. and Phillips, D.A. (1988). Nodulation and Symbiotic Dinitrogen Fixation. In *Alfalfa and Alfalfa Improvement*, 29, 229-257. (eds A.A. Hanson, D.K. Barnes and R.R. Hill). <https://doi.org/10.2134/agronmonogr29.c7>
- Volenec, J.J., Ourry, A., Joern, B.C. (1996). A role for nitrogen reserves in forage regrowth and stress tolerance. *Physiologia Plantarum* 97: 185–193.
- Wang, Z., & Şakiroğlu, M. (2021). The origin, evolution and genetic diversity of alfalfa. In L.-X. Yu, & C. Kole (Eds.), *The Alfalfa Genome* (pp. 29-31). Springer Cham.
- Zhao, D., Mejia-Guerra, K. M., Mollinari, M., Samac, D., Irish, B., Heller-Uszynska, K., ... & Sheehan, M. J. (2023, October). A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.). In *Genetic resources* (Vol. 4, No. 8, pp. 55-63).

Chapter 2

Analysis of Non-Dormant Alfalfa (*Medicago sativa*) Genetic Diversity and Population Structure using Genome Wide Microhaplotypes Markers

Kreingrai Nonkum, Matt Francis, Brian M. Irish, Scott Newell, Daniel Pap, Craig T. Beil,
Moir Sheehan, Dongyan Zhao, Meng Lin, and E. Charles Brummer

Abstract

Alfalfa (*Medicago sativa* L.) is one of the most commercially important forage crops in the US; however, yield improvement has stagnated, partly due to limited genetic diversity in commercial cultivars. Plant gene banks harbor large numbers of accessions containing useful alleles that could broaden the genetic base of non-dormant alfalfa adapted to the southern US and other non-dormant regions. The objective of this study was to investigate genetic diversity among 376 alfalfa accessions from seven geographic origins (Arabian Peninsula, India, Iran, Levant, Northwest Africa, Northeast Africa, and South America) using filtered DArTag microhaplotype markers. A panel of 3,000 SNP loci from the Diversity Array Technology (DArTag) platform, harboring 6,130 microhaplotypes, was used to quantify genetic diversity and population structure. The Bayesian model-based clustering program STRUCTURE was used to estimate the number of genetic clusters (K) among these accessions. Results microhaplotype data had greater polymorphism than target SNPs alone. Population structure analysis using The Evanno ΔK method identified five optimal clusters, corresponding to geographic regions: Levant, Turkey/Levant,

Africa, Iran & India, and South America. Principal component analysis (PCA) explained 7.7% and 6.9% of total variance with PC1 and PC2, respectively, with accessions from multiple geographic regions overlapping in the center of the PCA plot. The inbreeding coefficient (F_{IS}) was negative across all five clusters, ranging from -0.24 to -0.27 , indicating excess heterozygosity. Interpopulation genetic diversity, assessed by AMOVA, revealed high within-population variance and low among-population variance. These results demonstrated the utility and high reproducibility of the alfalfa 3K DArTag marker panel for genetic diversity studies in polyploid forage crops and show that microhaplotypes can reveal multiple alleles per accession with potential value for alfalfa breeding.

1. Introduction

Alfalfa (*Medicago sativa* L.), widely known as “the queen of forages,” is valued for its high productivity, nutritional quality as a livestock feed, and positive environmental benefits, including biological nitrogen fixation. In the United States, the estimated direct value of alfalfa in 2025 was \$7.6 billion, with a harvested area of 14.6 million acres (NASS, 2025), comparable to wheat and ranking first among forage crops (Putnam & Meccage, 2022). Cultivated alfalfa is an insect-pollinated, allogamous species that is autotetraploid ($2n=4x=32$), exhibiting tetrasomic inheritance (Bingham et al., 1994), with a haploid genome size of 816Mb (Chen et al., 2020). Due to its allogamous nature, alfalfa populations consist of highly heterozygous genotypes (Wilson et al., 2004). This allelic diversity enables adaptation to a wide range of environments under various abiotic and biotic stresses and provides the foundation for selection of agronomically important traits. Despite many breeding successes over the past century, particularly in pest and disease resistance, on-farm alfalfa biomass yield has remained stagnant, with little to no genetic gain over

the past 30 years (Brummer & Putnam, 2018). Although the causes of this yield stagnation are not fully understood, developing novel breeding strategies could accelerate yield gains and improvement of other quantitative traits, including tolerance to salinity, cold, heat, and drought.

Alfalfa cultivar dormancy classes are determined by the amount of regrowth produced in the fall following a late-season harvest, with fall dormancy (FD) scores ranging from 1 (very dormant) to 11 (extremely nondormant) (Teuber et al., 1998). Varieties adapted to northern USA and Canada are typically classified as dormant or moderately dormant (FD = 2–5), exhibiting high winter hardiness but producing lower yields than those produced regions suited for year-round production. Less dormant cultivars, while susceptible to winter injury in cold climates, can grow throughout the winter in warm regions such as Florida, southern California, and Arizona. Non-dormant alfalfa varieties were generally developed from germplasm sources genetically distinct from those used in midwestern and northeastern US cultivars (Şakiroğlu & Brummer, 2007). Hybridizing dormant and non-dormant varieties has shown yield improvement potential in some cases (Brummer 1999; Segovia-Lerma et al., 2004), though no improvement was observed in others (Şakiroğlu & Brummer, 2007).

Genetic diversity is fundamental to crop improvement. Insufficient diversity limits genetic gain in breeding programs and constrains adaptation to changing environments, a concern of growing urgency in the era of rapid climate change (Govindaraj et al., 2015). Germplasm collections maintained in gene banks worldwide represent a valuable reservoir of diversity that can be incorporated into breeding programs to broaden the genetic base available for selection (Simmonds, 1993).

The National Plant Germplasm System, maintained by the USDA Agricultural Research Service, preserves plant germplasm resources accessible through the GRIN-Global database (<https://www.grin-global.org/>) (Byrne et al., 2018; Postman et al., 2006). For alfalfa, the temperate-adapted forage legume collection contains approximately 4,000 accessions of cultivated

alfalfa and its wild relatives (Irish & Greene, 2021). Given the scale of this collection, efficient tools for characterizing and identifying accessions are essential for leveraging this diversity in alfalfa improvement programs.

Genetic markers are widely used to assess genetic diversity and identify loci associated with traits of interest. A Diversity Array Technology (DArTag) genotyping platform was recently developed for alfalfa by Breeding Insight (Zhao et al., 2023; Medina et al., 2024; Zhao et al., 2024). This platform interrogates 3,000 SNP loci distributed across the alfalfa genome. At each locus, oligonucleotide probes flanking a 54 base pairs (or 81 base pairs in the current version) region around the target SNP capture sequence data that enable identification of both target and off-target SNPs. The combination of alleles at the target SNP and any nearby off-target SNPs within the captured region is referred to as a microhaplotype (Zhao et al., 2023; Medina et al., 2024). By considering all polymorphisms within each captured region, rather than the target SNP alone, multiple haplotype alleles per locus can be identified. In an autotetraploid species such as alfalfa, up to four alleles may be present per individual, making microhaplotypes particularly informative for characterizing polyploid genetic diversity.

In this study, we investigated the genetic diversity of 376 nondormant alfalfa accessions from the national collection with potential utility for non-dormant alfalfa breeding programs in the US. Despite the promise of the DArTag platform, its utility for characterizing genetic diversity and population structure in nondormant alfalfa accessions spanning diverse geographic origins had not previously been evaluated. We hypothesized that genetic variation among these accessions would reflect geographic origin, with greater differentiation between geographically distant regions than among accessions from nearby regions. The objective of this study was to (1) evaluate the genetic

diversity and population structure of these nondormant alfalfa accessions using the alfalfa DArTag microhaplotype marker panel.

2. Materials and Methods

2.1 Plant Materials

A total of 376 nondormant alfalfa accessions were obtained from the National Plant Germplasm System via GRIN (<https://www.ars-grin.gov/>). The accessions represented alfalfa collected from regions with warm Mediterranean climates and moderate winter temperatures, including the Arabian Peninsula, India, Iran, the Levant (the eastern Mediterranean coastal region), northwest Africa, northeast Africa, and South America. Fall dormancy check cultivars (Teuber et al., 1988) spanning FD = 4–11 and three check entries (WL 365HQ, AmeriStand 803T, and UF_AlfPers_2015, an experimental line from the University of Florida) were included for comparison (Supplementary File 1).

2.2 DNA extraction

Seeds of each accession were germinated in the greenhouse at the National Temperate Forage Legume Genetic Resources Unit in Prosser, WA. A single trifoliolate leaf (~50 mg fresh weight) was collected from each plant. Twenty plants per accession were sampled for DNA extraction, with ten plants allocated to each of two replicate plates. DNA was extracted using DNeasy 96 Plant Kits (Qiagen, catalog no. 69181). Following extraction, aliquots from the two replicate samples were pooled into a third plate, so that each accession was represented by three DNA samples: two independent 10-plant pools and one 20-plant bulk pool derived from combining both 10-plant pools.

2.3 DArTag genotyping

Genotyping and data filtering were conducted as described by Medina et al. (2025). Briefly, DNA samples were shipped to Diversity Array Technologies (DArT; www.diversityarrays.com, Australia) for DArTag genotyping. Because DArTag is an amplicon-based targeted genotyping platform, it captures and sequences the region flanking each target SNP so that non-target SNPs may be identified. The combination of alleles at target and off-target loci within the captured region is referred to as a microhaplotype. Sequencing read depths for both target and off-target SNP within the 3000 genomic regions are provided by DArT in the Missing Allele Discovery Count (MADC) file. The nucleotide at the target SNP is denoted as either “reference” or “alternative” microhaplotype. Microhaplotypes with off-target SNP that contain the reference target allele are denoted as RefMatch microhaplotypes and those containing the alternate target allele are AltMatch microhaplotypes.

2.4 Filtering and allele genotype calling

Marker data were analyzed as genotypes rather than as allele frequencies to facilitate downstream analysis. Loci were retained for genotyping only if they had at least 10 sequencing reads in $\geq 95\%$ of samples and all microhaplotypes were retained for genotyping only if they had at least 10 sequencing reads in $\geq 95\%$ of samples; genotypes were then assigned at each retained locus from the observed microhaplotypes. The filtered MADC file indicating the read count for the target SNP genomic regions meeting the filtering criteria was imported into the PolyRAD v.2.0.0 R package (Clark et al., 2019) using the function ‘readDArTag’. RADdata objects were used as an input file for genotype estimation using the ‘IteratePopStructure’ function in PolyRAD. The final genotype calls at the microhaplotype level were exported as a Variant Call Format (VCF) file using the ‘RADdata2VCF’ function with option `asSNPs = False`. Allele dosage of the genotypes the 3K target SNPs was called following similar steps as those performed for microhaplotype but included only reference and

alternative microhaplotypes from the filtered MADC. The SNP density plot was generated using Cmplot website-based application (Tang et al., 2023)

2.5 Population structure inferring

Bayesian population structure was inferred using STRUCTURE software (Pritchard et al., 2000). A model-based approach was used to identify K clusters within the marker dataset. Simulations were run using an admixture model with correlated allele frequencies and tetraploid genotypic values. Parameters were set to 10,000 burn-in iterations followed by 10,000 Markov Chain Monte Carlo (MCMC) replicates.

The optimal number of clusters (K) was determined by testing $K = 1$ to $K = 10$ using the Evanno method (Evanno et al., 2005). This method was implemented via the ‘evannoMethodStructure’ function in ‘pophelper’ v.2.3.1 R package (Francis, 2017). Following identification of the best K value, population proportions were visualized using the ‘plotQ’ function.

2.6 Population structure

Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) were performed for population structure analysis. PCA was applied to target SNPs and microhaplotype markers using GenoDive v.3.0 (Meirmans, 2020). DAPC was restricted to target SNPs to meet the biallelic SNP requirement of the adegenet R package v2.1.10 (Jombart et al., 2010). The marker matrix, marker positions, plant IDs, and geographic regions were converted into a ‘genlight’ object, and the ‘dapc’ function was used to characterize genetic structure among the seven germplasm regions in this study.

2.7 Euclidian distance and neighbor-joining tree

An individual Euclidean distance matrix was calculated based on target SNP genotypes using GenoDive v.3.0 (Meirmans, 2020). The Euclidean distance matrix (d_{ij}) was calculated as

$$d_{ij}^2 = \sum_{a=1}^A p_{ia} - p_{ja}^2$$

Where p_{ia} is the frequency of a th allele ($a = \{1, 2, \dots, A\}$) within individual i , p_{ja} is the frequency of a th allele ($a = \{1, 2, \dots, A\}$) within individual j ; in autotetraploids, these frequencies can take values of 0, 0.25, 0.5, 0.75, or 1 (Meirmans & Liu, 2018).

A neighbor-joining tree was then constructed from the Euclidean distance matrix using the ‘ape’ R package (Paradis & Schliep, 2019) to illustrate genetic relationships among the accessions.

The VCF file with only biallelic target SNP was imported to create RADobject using ‘VCF2RADdata’ function, the RADdata object was converted to the Structure format file using ‘Export_Structure’ using PolyRAD.

2.8 Genetic diversity

Genetic diversity indices were calculated for target SNPs and microhaplotype markers, based on both individual samples and STRUCTURE clusters, using GenoDive v.3.0 (Meirmans, 2020). Intrapopulation genetic diversity metrics were percentage of monomorphic loci (%mon), number of alleles (Num), effective number of alleles (N_E), observed heterozygosity (H_o), expected

within-population heterozygosity (H_s), inbreeding coefficient (F_{IS}), and polymorphism information content (PIC), calculated using R software.

Analysis of Molecular Variance (AMOVA) was used as interpopulation genetic diversity. To evaluate genetic marker diversity among and within populations, we used AMOVA adapted to autopolyploids (Meirmans & Liu, 2018). AMOVA was calculated using the software GenoDive v.3.0.0 (Meirmans, 2020).

3. Results

3.1 DArTag Marker distribution

After applying the filtering criteria, 2,769 of the 3,000 SNP marker loci with less than 5% missing data across all accessions were retained for analysis. From these loci, 6,130 microhaplotypes were identified, providing a more robust framework for assessing genetic diversity than the 2,769 target SNP markers alone. The retained markers were distributed across all eight chromosomes with an average density of 4.42 SNPs/Mb (Table 1). SNP density varied by chromosome, ranging between 1.49 SNPs/Mb on Chromosome 6, which exhibited large gaps in the centromeric region, to 5.06 SNPs/Mb on Chromosome 1 (Figure 1).

Table 1 Distribution and density (markers per megabase, Mb) of the filtered DArTag markers anchored across the alfalfa reference genome assembly, based on 376 non-dormant entries.

Chromosome	Marker Number	Genome size (Mb)	SNP/Mb
Chr1	414	81.72	5.06
Chr2	361	76.26	4.73
Chr3	416	92.27	4.50
Chr4	426	89.46	4.76
Chr5	389	80.92	4.80
Chr6	209	79.94	2.61
Chr7	365	88.21	4.13
Chr8	189	37.54	5.03
Total	2769	626.34	4.42

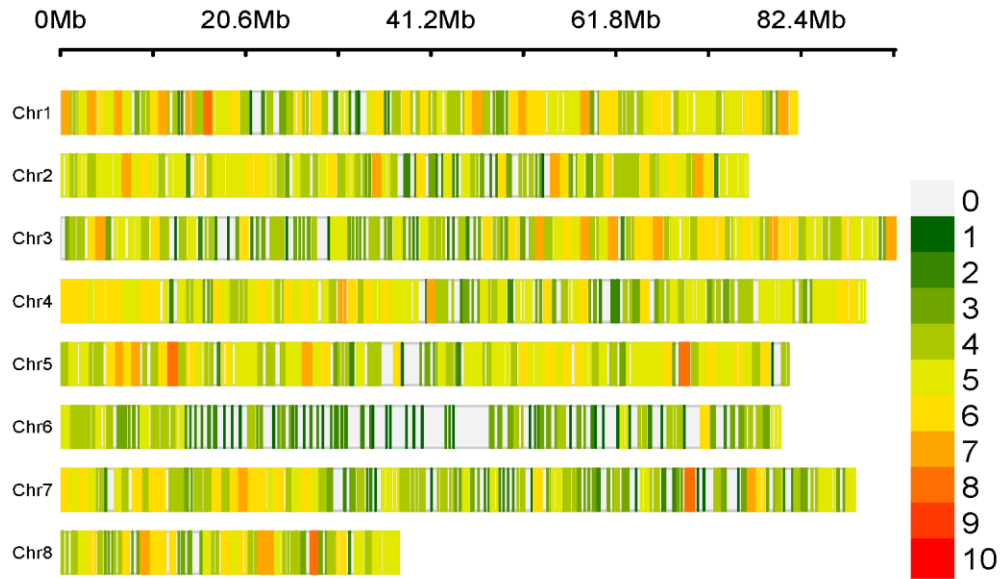


Figure 1. SNP density across the eight chromosomes of the alfalfa genome for the filtered 2,769-marker DArTag panel.

3.2 Marker polymorphism

Across all 376 accessions, 33% of target SNP loci but only 4% of microhaplotype loci were monomorphic, confirming the greater polymorphic resolution of microhaplotypes relative to target SNPs (Table 3). Among the five STRUCTURE clusters (section 2.5), Cluster 2 had the highest average percentage of monomorphic loci (14.7%) and Cluster 4 had the lowest (9.3%).

3.3 Microhaplotype summary

In this study, approximately 6,130 microhaplotypes from 2,769 DArTag loci were used to characterize the genetic diversity of nondormant alfalfa. The microhaplotypes identified here are approximately half the number (~12,000) reported in studies of dormant and pre-breeding alfalfa germplasm (Medina et al., 2024; Zhao et al., 2024). During the generation of the DArTag library for 376 non-dormant alfalfa accessions, the 2,769 markers were retained with Minor allele frequency (MAF) > 0.2.

Compared with target SNPs alone, microhaplotype markers substantially increased the informative content of the DArTag panel. Across 376 nondormant alfalfa accessions, microhaplotypes increased the total number of alleles in the study from 4,476 to 5,547 after filtering and microhaplotypes can reduce the percentage of monomorphic markers from 33% to 4%

3.4 Reproducibility and correlation

To validate the pooling strategy described in section 2.2, genotyping reproducibility was assessed by calculating correlation coefficients (r^2) of allele frequencies across each locus among the three replicated samples per accession, the two independent 10 plant bulks (Pool1; P1 and Pool2; P2) and the 20-plant bulk formed from the two 10 plant samples (Bulk). A strong overall correlation was observed across all comparisons with an average $r^2 = 0.95$, indicating high consistency in the allelic profiles generated by the DArTag genotyping platform (Figure 2).

As expected, the highest correlations were observed when comparing each individual pool against the bulk: P1 vs. Bulk ($r^2 = 0.97$), followed by Pool2 and Bulk (P2&Bulk) with $r^2 = 0.96$. The higher values are expected due to bulk includes the DNA from both Pool1 and Pool2. The most robust measure of biological repeatability is the comparison between Pool1 and Pool2, which yields a mean correlation of 0.92, with values ranging between 0.68 to 0.97 (Table 2, Figure 2A). The similarity of allelic profiles of two individual pools demonstrates the reliability of the pooling strategy for capturing the allelic profiles by 3K alfalfa DArTag.

Table 2. Summary Statistics of correlation of allele frequencies between three types of bulk, Pool1 (P1), Pool2, and Bulk (Pool1 + Pool2)

Comparison type	Mean correlation	Std Dev	Min	Max
P1&P2	0.92	0.01	0.67	0.96
P1&Bulk	0.97	0.01	0.91	0.98
P2&Bulk	0.96	0.01	0.71	0.98

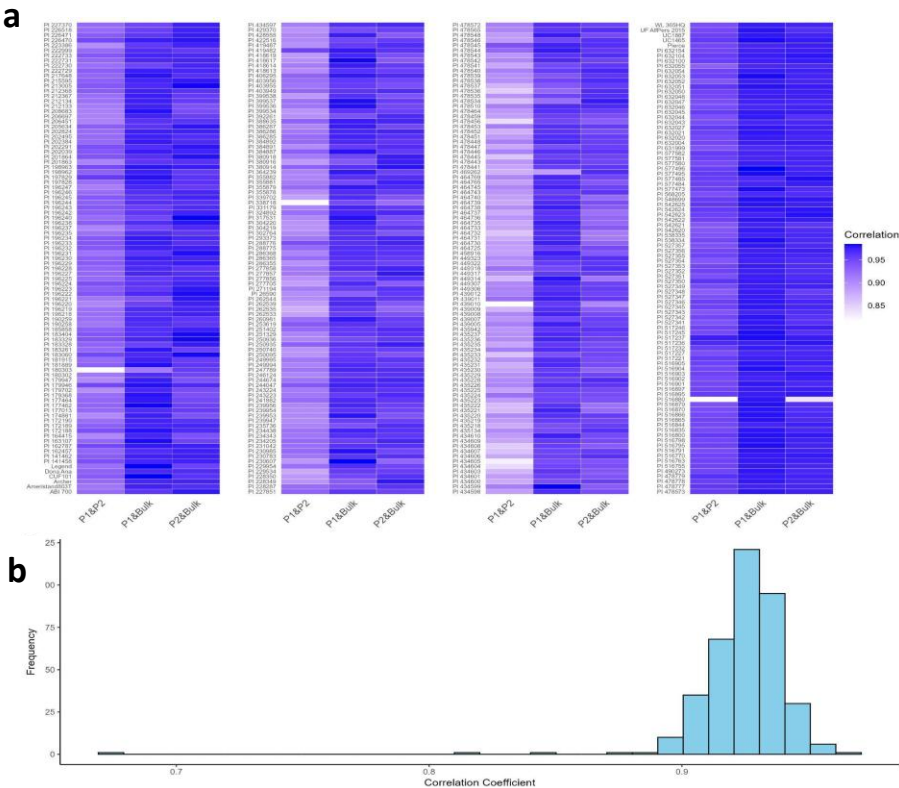


Figure 2. Correlation of allele frequency per marker of pool 1 (P1), pool 2 (P2) and bulk of pool1 and pool2 (Bulk) for each 376 nondormant alfalfa accessions. (a) heatmap of correlation of P1&P2 P1&Bulk and P2&bulk for each 376 nondormant alfalfa accessions. (b) Histogram of correlation coefficient among P1 and P2 samples for each 376 nondormant alfalfa accessions.

3.5 Inter-population genetic diversity parameters

Consistent with their greater polymorphism, microhaplotypes showed higher values than target SNPs for all intra-population diversity metrics: number of alleles (Num), effective number of alleles (N_E), observed heterozygosity (H_O), and expected heterozygosity within population (H_S) of microhaplotypes were higher than of target SNPs. (Table 1).

The inbreeding coefficient (F_{IS}) reflects deviation from random mating: values below zero indicate excess heterozygosity (outbreeding), while values above zero indicate excess homozygosity (inbreeding).

F_{IS} mean values were slightly lower for microhaplotypes (-0.25) than for target SNPs (-0.27). F_{IS} values were negative for both marker types, indicating high heterozygosity across the panel, as is typical for the allogamous alfalfa. The highest F_{IS} was observed in Cluster 2 (-0.24) and the lowest in Cluster 4 (-0.27) (Table 3).

Table 3. Intra-population genetic diversity of target SNPs and microhaplotypes of five cluster using STRUCTURE for 376 nondormant alfalfa accessions which are percentage of monomorphic loci (% Mon), number of alleles (Num), effective number of allele (NE), observed heterozygosity (H_o), sub population heterozygosity (H_s), inbreeding coefficient (F_{is}) and Polymorphism information content (PIC)

Pop	Size	Mono. Loci (%)	No. Alleles	Effective No. Alleles (Ne)	Obs. Het. (H_o)	Subpop. Het. (H_s)	F_{is}	PIC
Target SNPs								
Cluster1	35	27.50	1.69	1.48	0.34	0.27	-0.26	0.17
Cluster2	40	33.70	1.68	1.45	0.32	0.26	-0.25	0.13
Cluster3	110	31.17	1.67	1.45	0.33	0.25	-0.28	0.11
Cluster4	99	33.41	1.66	1.46	0.33	0.26	-0.29	0.18
Cluster5	92	32.47	1.67	1.45	0.33	0.26	-0.28	0.13
All	376	32.97	1.73	1.46	0.33	0.26	-0.27	-
Microhaplotypes								
Cluster1	35	13.94	1.90	1.55	0.38	0.31	-0.25	0.24
Cluster2	40	14.73	1.89	1.51	0.36	0.29	-0.24	0.22
Cluster3	110	9.31	1.95	1.51	0.36	0.29	-0.26	0.2
Cluster4	99	9.28	1.95	1.52	0.37	0.29	-0.27	0.26
Cluster5	92	9.46	1.94	1.51	0.36	0.29	-0.26	0.21
All	376	4.11	2.00	1.52	0.37	0.3	-0.25	-

3.6 STRUCTURE population inference

Using the Evanno method implemented in ‘pophelper’ (section 2.5), the optimal number of subpopulations was determined to be $K = 5$, with a secondary peak at $K = 3$ (Supplementary Figure 1). At $K = 2$, The Bayesian model-based clustering program, STRUCTURE separated the Levant group from all other geographic origins. At $K = 3$, the plot shows three distinct clusters of Levant, Africa, and South America. The result indicates that the two major centers of origin of alfalfa, the first center rooted in Levant particularly within Turkey and Iran and a second center located in Africa and South America, which can be traced back to origins in Spain and Central Asia. (Wang & Şakiroğlu, 2021; Russelle, 2001). When the $K=5$ in the model, indicated by the Evanno analysis (Supplementary Figure 1.), the clusters align more closely with specific geographical distribution (Figure 3).

Each cluster was named according to the geographic origin most frequently represented in its membership: Levant (Cluster 1), Iran and Turkey (Cluster 2), Arabian Peninsula and Africa (Cluster 3), Iran, India, and Africa (Cluster 4), and South America (Cluster 5) (Figure 4).

Table 4. Cluster membership based on STRUCTURE analysis (K = 5)

Cluster	Total	Geographic origins (n)	Check varieties (FD)
1	35	Levant (15), South America (6), North West Africa (6), Iran (2), India (2)	WL365HQ (FD 4.9), Legend (FD 4), Archer (FD 5), ABI700 (FD 6) (4 check varieties)
2	40	Levant (20), Iran (20)	None
3	110	Arabian Peninsula (36), North East Africa (36), North West Africa (16), India (11), Levant (6), South America (2)	CUF101 (FD 9), UC 1465 (FD 11), UC 1887 (FD 10) (3 check varieties)
4	99	India (39), Iran (22), North West Africa (22), Levant (10), North East Africa (3)	Ameristand 803T (FD 8), Dona Ana (FD 7), Pierce (FD 8) (3 check varieties)
5	92	South America (79), Levant (5), North West Africa (4), North East Africa (2), India (1)	UF AlfPers 2015 (FD 9) (1 check variety)

FD = Fall Dormancy rating. Accession counts by geographic origin shown in parentheses.

Cluster membership at $K = 5$ is summarized in Table 4. Cluster 1 comprised 35 accessions from Levant ($n = 15$), South America ($n = 6$), Northwest Africa ($n = 6$), Iran ($n = 2$), and India ($n = 2$), plus 4 check cultivars (WL365HQ, FD = 4.9; Legend, FD = 4; Archer, FD = 5; ABI700, FD = 6).

Cluster 2 contained 40 accessions exclusively from Levant ($n = 20$) and Iran ($n = 20$), with no check cultivars.

Cluster 3 was the largest, comprising 110 accessions from the Arabian Peninsula ($n = 36$), Northeast Africa ($n = 36$), Northwest Africa ($n = 16$), India ($n = 11$), Levant ($n = 6$), and South America ($n = 2$), along with 3 check cultivars (CUF101, FD = 9; UC 1465, FD = 11; UC 1887, FD = 10).

Cluster 4 contained 99 accessions from India ($n = 39$), Iran ($n = 22$), Northwest Africa ($n = 22$), Levant ($n = 10$), and Northeast Africa ($n = 3$), with 3 check cultivars (AmeriStand 803T, FD = 8; Dona Ana, FD = 7; Pierce, FD = 8).

Cluster 5 included 92 accessions, predominantly from South America ($n = 79$), with additional accessions from Levant ($n = 5$), Northwest Africa ($n = 4$), Northeast Africa ($n = 2$), and India ($n = 1$), and 1 check cultivar (UF AlfPers 2015, FD = 9).

Notably, the check cultivars clustered in a manner consistent with their fall dormancy scores: Clusters 1 contained semi-dormant cultivars (FD < 6), Cluster 4 contained non-dormant cultivars (FD 7–8), and Clusters 3 and 5 represented very non-dormant accessions (FD 9–11). The Florida-bred FD = 9 cultivar (UF AlfPers 2015) grouped with the predominantly South American Cluster 5 (Figure 5), consistent with the known germplasm background of this experimental line.

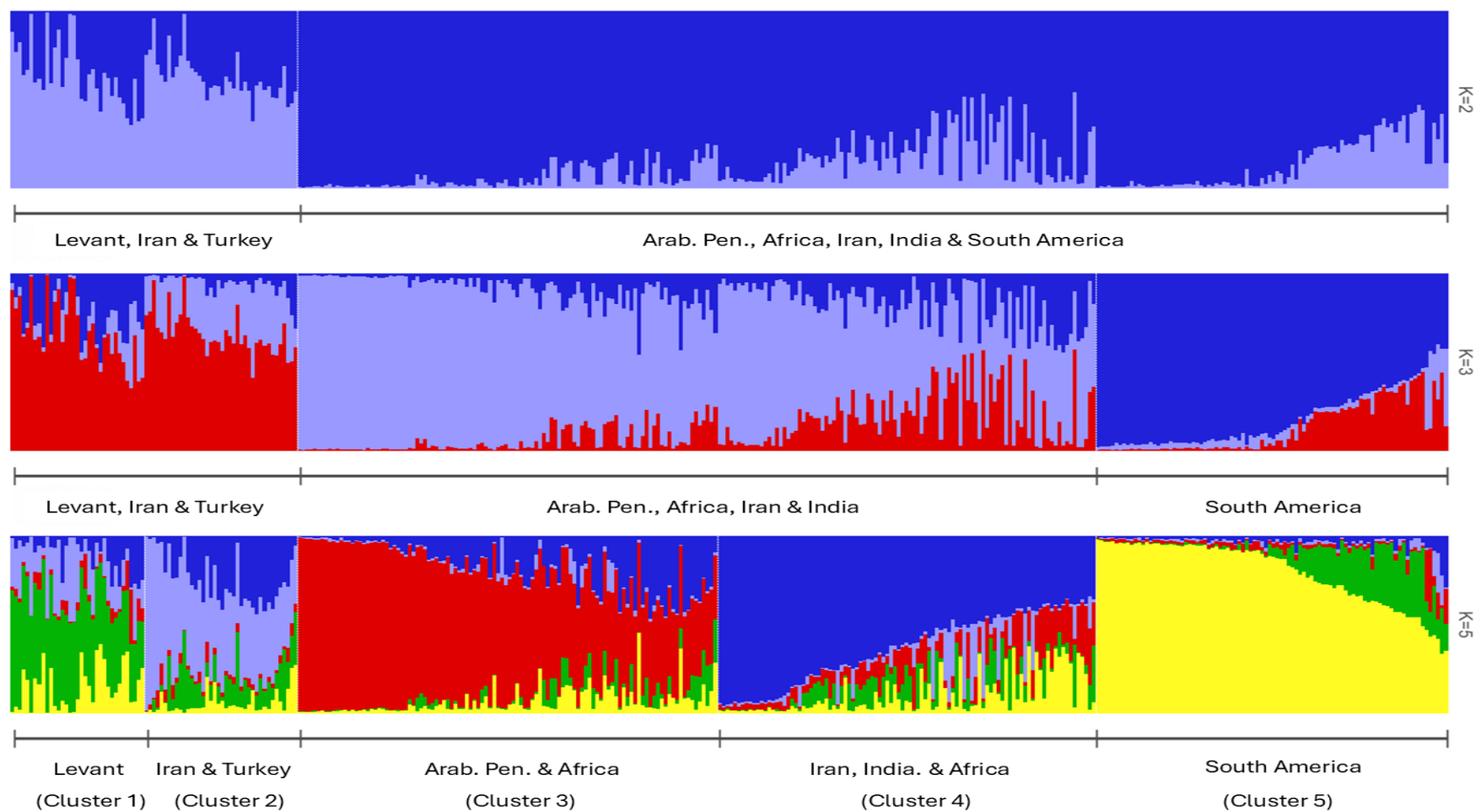


Figure 3. Nondormant alfalfa population structure based on Bayesian inference (STRUCTURE) among 376 alfalfa accessions assuming K=2, K=3 and K=5 clusters.

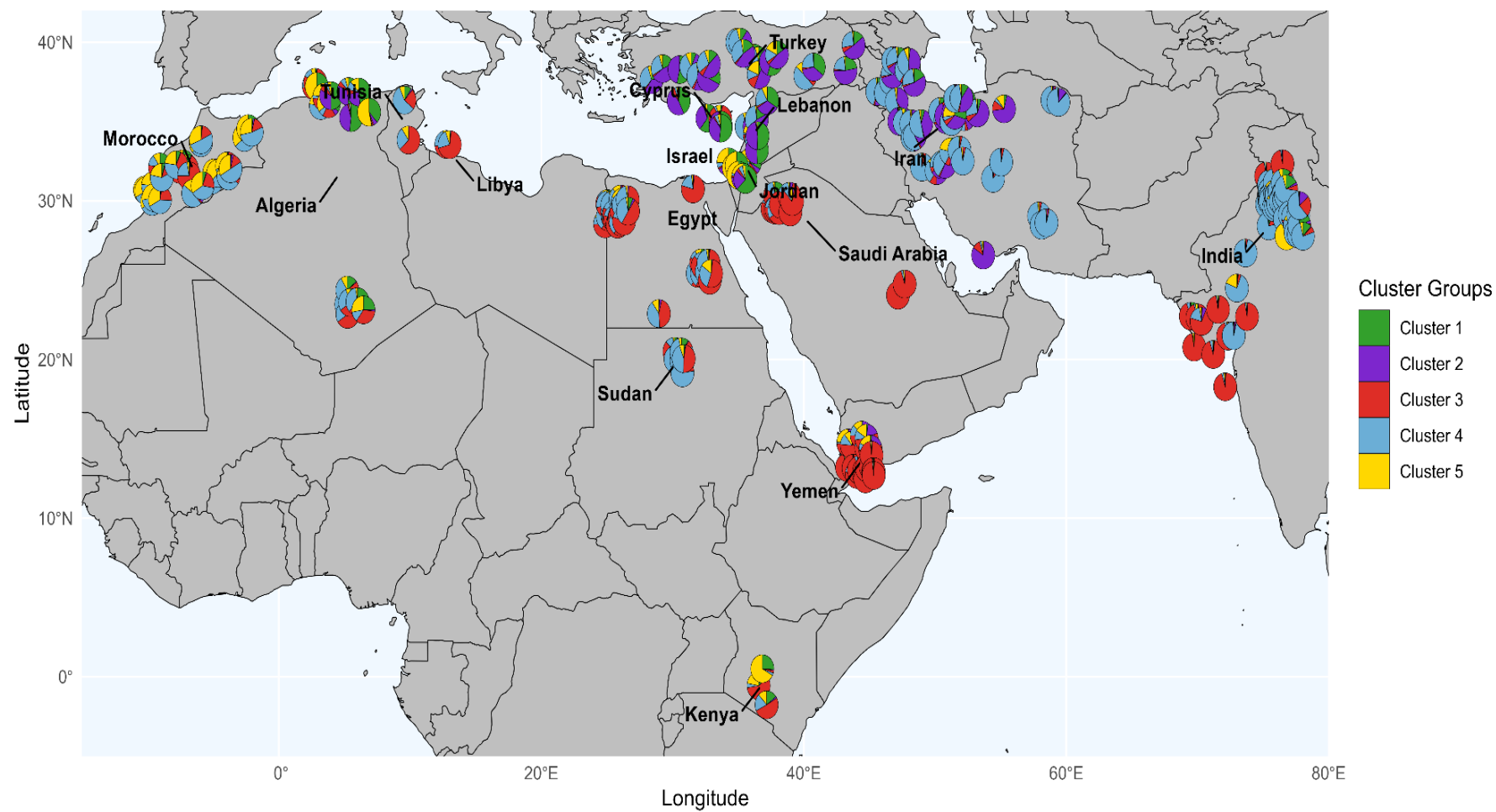


Figure 4. Map indicating the collection location for 376 accessions used in the study. The pie chart shows $K=5$ structure portions for the accessions collected from India, Levant and Africa.

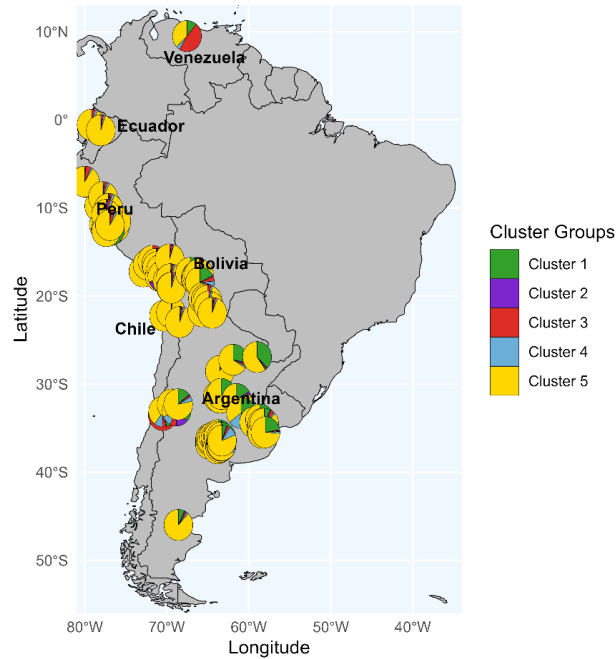


Figure 5. Map indicating the collection location for 376 accessions used in the study. The pie chart shows K =5 structure portions for the accessions collected from South America.

3.7 Principal component analysis

PCA was performed on the 376 accessions using the 2,769 target SNPs (section 2.6). PC1 and PC2 explained 7.7% and 6.9% of total variance, respectively. Although these values reflect the high dimensionality of diversity across the panel, the PCA plot revealed clear geographic structuring for several groups (Figure 6). South American accessions formed a discrete cluster at the bottom of the plot, while Iranian and Levant accessions occupied the upper-left region. Accessions from Northeast Africa, Northwest Africa, the Arabian Peninsula, and India overlapped broadly in the upper-right region of the plot, consistent with their shared geographic and genetic backgrounds. Check cultivars clustered centrally, reflecting their intermediate genetic composition. A small number of accessions plotted outside their expected geographic regions, likely reflecting admixed or misclassified germplasm.

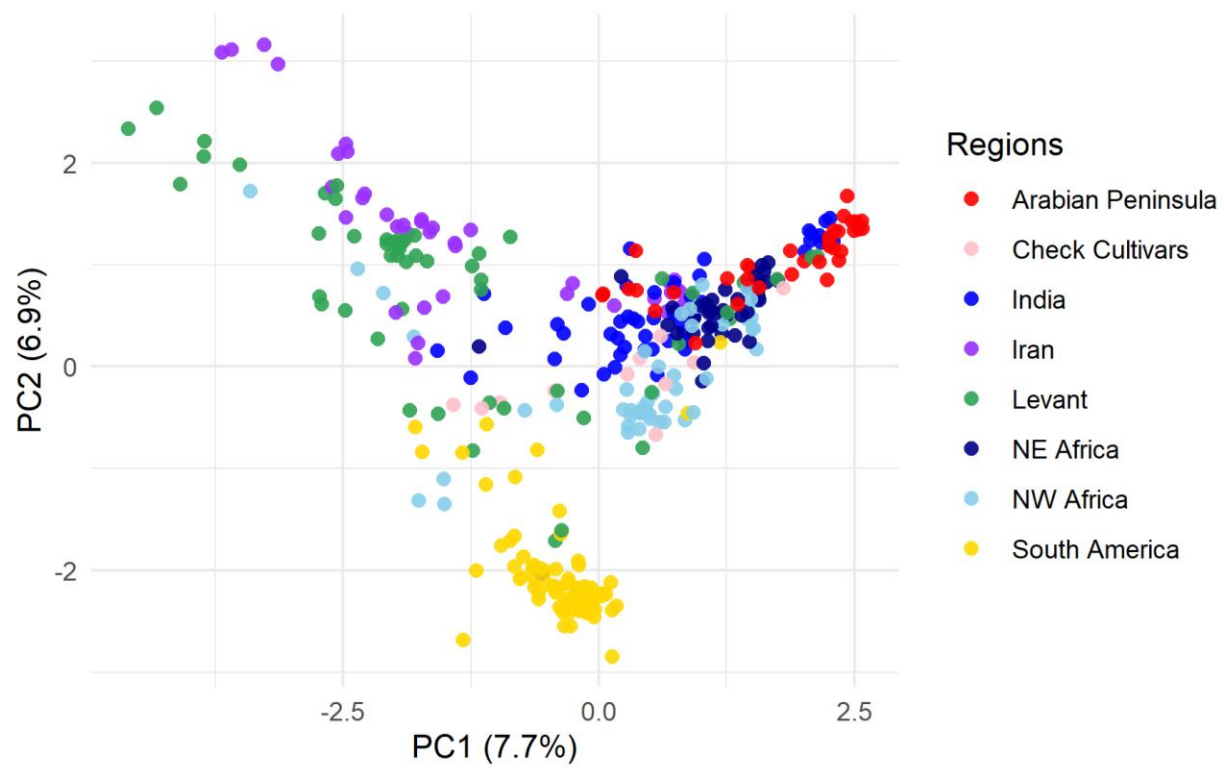


Figure 6. Plot of PC1 and PC2 of 376 nondormant alfalfa germplasm contains accessions from Arabian Peninsula, Check Cultivars, India, Iran, Levant (Mideast), Northeast Africa (NE Africa), Northwest (Africa) and South America.

3.8 DAPC analysis of population structure

DAPC was applied to the target SNP dataset to complement PCA with a supervised, boundary-maximizing analysis. The results revealed a clearer delineation of genetic groups than standard PCA. The South American accessions were the most genetically divergent, forming a well-separated cluster distinct from all others. The Iranian accessions also displayed high genetic distinctiveness, forming a discrete cluster at the bottom left of the plot (Figure 7).

Accessions from the Arabian Peninsula and Northeast Africa showed considerable overlap in the DAPC plot, consistent with historical gene flow across the Red Sea and the Sahara Desert.

The Levant, Northwest Africa, and India accessions showed the greatest degree of overlap, suggesting shared genetic background across these three regions, likely reflecting historical trade and germplasm exchange along ancient routes.

Together, PCA and DAPC provided complementary views of the genetic structure of the nondormant germplasm. PCA captured the continuous distribution of diversity, with South American and Levant/Iranian germplasm at opposite extremes and a central admixture zone where accessions from Arabian Peninsula, NE/NW Africa, Levant, and India converged (Figure 7).

The DAPC confirms the patterns found in the PCA plot by forming the boundaries between specific geographical regions. The DAPC plot reinforces the isolated, highly diverse group of the South America cluster and the Iranian cluster from the other cluster.

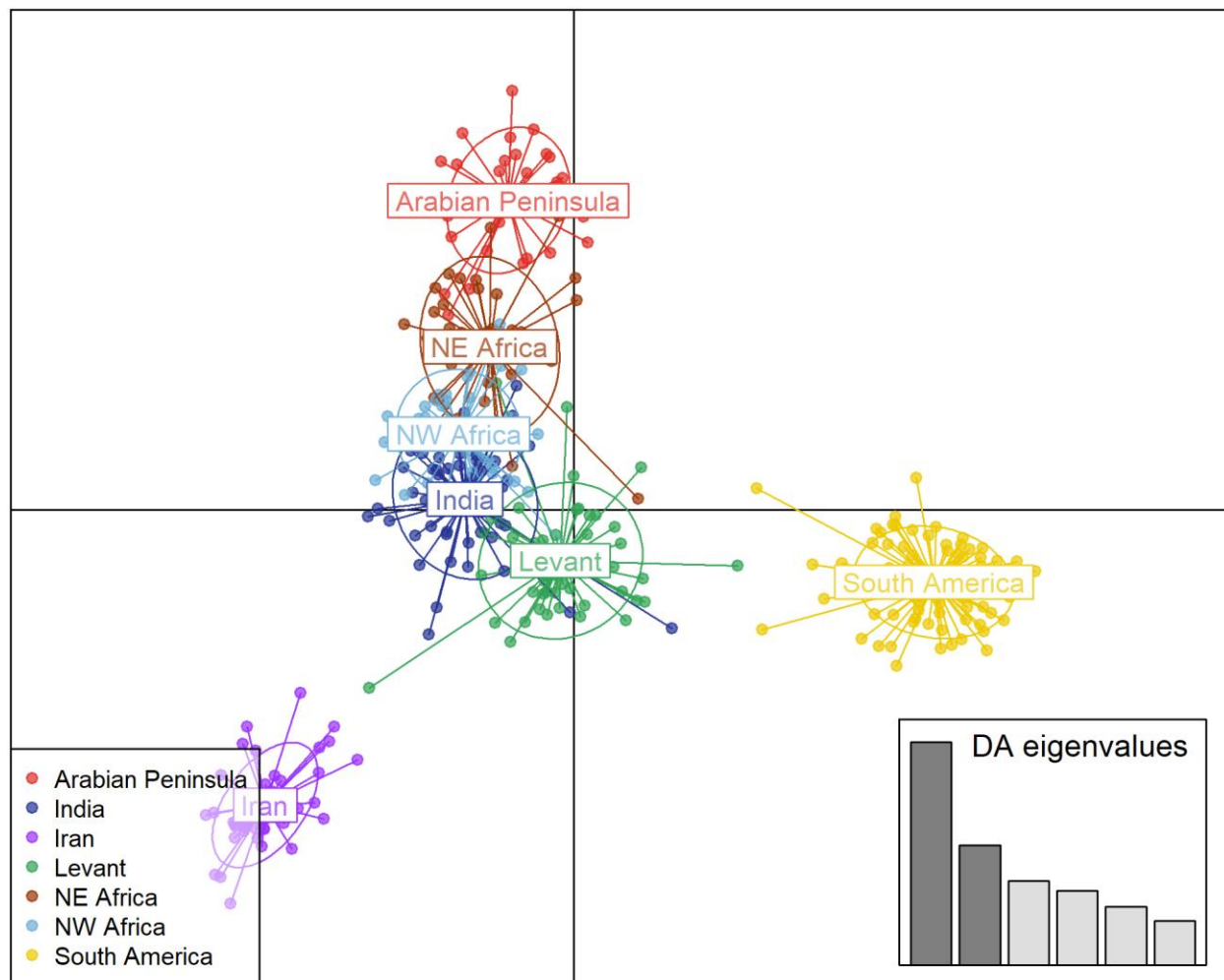


Figure 7. DAPC scatter plot of genetic structure of 376 non-dormant alfalfa germplasm

3.9 Neighbor-Joining (NJ) Phylogenetic Analysis

To complement the population structure analyses, a Neighbor-Joining (NJ) tree was constructed from pairwise Euclidean distances calculated from the 2,769 target SNPs. The resulting relationship tree broadly corresponded to geographic origins, further corroborating the structure identified by STRUCTURE, PCA, and DAPC (Figure 8).

Consistent with previous analyses, South American accessions formed a highly distinct, well-supported clade with pronounced divergence from all other geographic groups. Iranian accessions formed a discrete subset with a closer relationship to Levantine and Indian clusters than to African or Arabian Peninsula accessions.

In contrast, accessions from the Arabian Peninsula, NE Africa, NW Africa, and a subset of India exhibited a more complex, intermixed branching pattern, consistent with a high degree of historical germplasm exchange across these regions. The close intermixing of Arabian Peninsula and NE Africa accessions supports the hypothesis of gene flow across the Red Sea corridor. Overall, the NJ tree illustrated the broad geographic distribution of nondormant alfalfa germplasm across Eurasia and Africa, and the genetic isolation of the South American lineage.

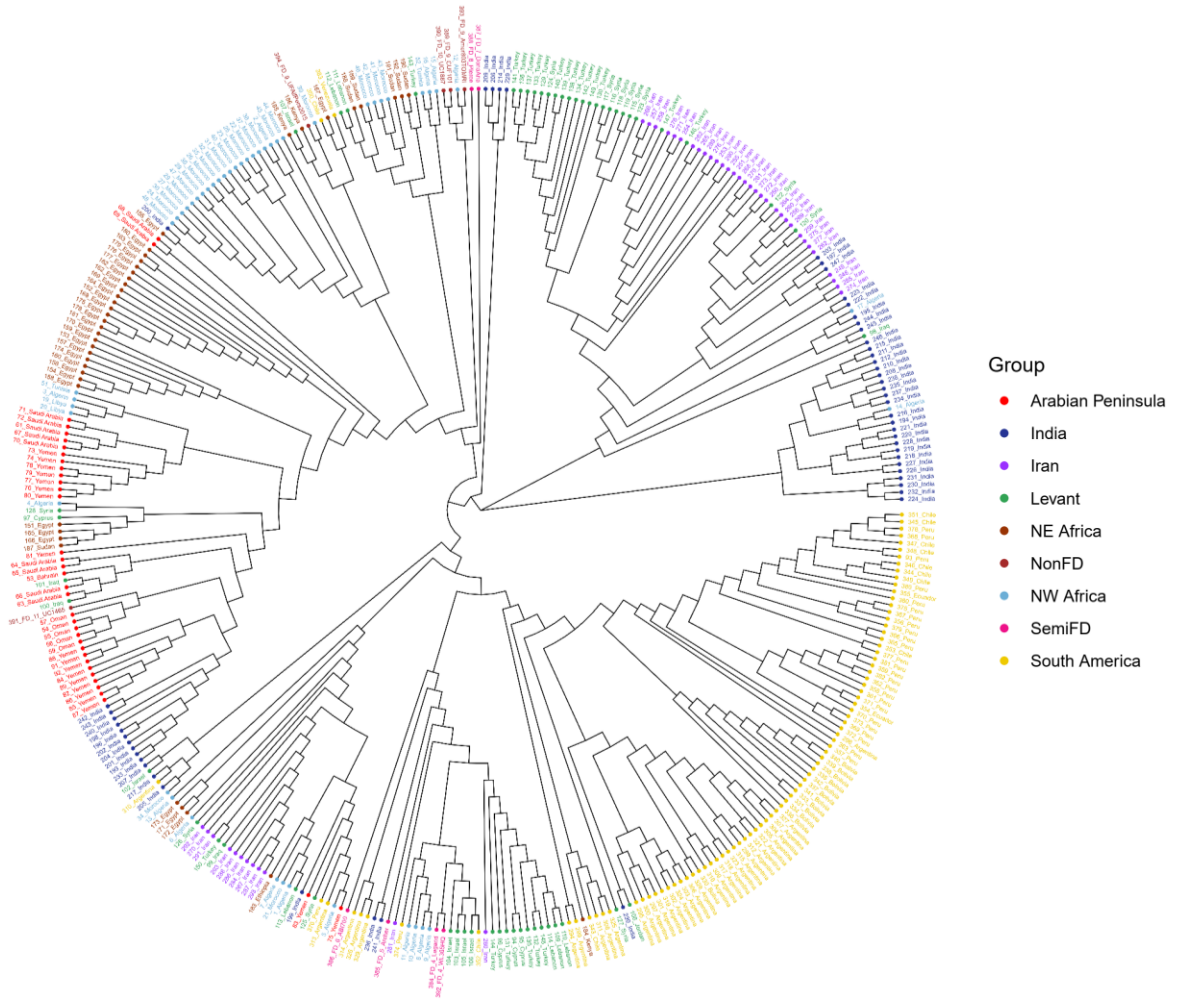


Figure 8. Neighbor-Joining phylogenetic tree based on pairwise Euclidean distances using 2,769 target SNPs

3.10 The Analysis of Molecular Variance (AMOVA)

AMOVA was performed to partition total genetic variance within and among the five STRUCTURE clusters (section 2.8). The five clusters explained 15% and 18% of total genetic variance for target SNPs and microhaplotypes, respectively, while within-population variance accounted for 85% and 82% of total variance. The among-population component was highly significant for both marker types ($P < 0.01$) based on 1,000 random permutations, confirming that the five clusters represent genetically differentiated groups despite the predominance of within-population diversity (Table 5).

Table 5. Analysis of Molecular variance (AMOVA) of five STRUCTURE clusters of 376 nondormant alfalfa population

Source of Variation	SSD	d.f.	MS	Var-comp	%Var	P-value
Target SNPs						
Within Population	11260.17	371	30.35	30.35	0.85	--
Among Population	1663.71	4	415.93	5.36	0.15	0.001
Microhaplotypes						
Within Population	16435.61	371	44.30	44.30	0.82	--
Among Population	2974.28	4	743.57	9.72	0.18	0.001

4. Discussion

4.1 DArTag Microhaplotypes: Cross-Study Comparability and Marker Efficiency

In this study, approximately 6,130 microhaplotypes from 2,769 DArTag loci were used to characterize the genetic diversity of nondormant alfalfa. Zhao et al. (2023) demonstrated the potential of the MADC file within the DArTag platform for identifying additional alleles beyond the target SNP, and the present results build on this framework. A key advantage of microhaplotypes is that each locus ID is consistent across DArTag studies, enabling direct cross-study comparisons and validation of alleles at specific genomic positions (Zhao et al., 2023; Zhao et al., 2024; Medina et al., 2024). The microhaplotypes identified here are approximately half the number (~12,000) reported in studies of dormant and pre-breeding alfalfa germplasm (Medina et al., 2024; Zhao et al., 2024), likely reflecting the narrower genetic diversity of nondormant accessions and underscoring the value of expanding diversity in nondormant breeding programs. Microhaplotypes can reduce the percentage of monomorphic markers from 33% to 4%; this result is supported by the DArTag marker study on US alfalfa germplasm by Medina et al. (2024).

4.2 National Plant Germplasm and the DArTag Platform for Diversity Assessment

The foundation of this research is that the National Plant Germplasm System and similar repositories worldwide harbor alfalfa accessions carrying novel alleles useful for cultivar improvement (Simmonds, 1993). By applying the DArTag targeted sequencing platform to 376 nondormant alfalfa accessions, this study assessed the genetic diversity and geographic structure of germplasm potentially valuable for breeding programs in non-dormant regions such as California and the southern US. Understanding these genetic relationships enables breeders to identify complementary accessions, plan strategic crosses, and prioritize germplasm for

evaluation. The microhaplotype approach further enhanced this capacity by detecting multiple alleles per locus within short sequenced regions, improving diversity resolution compared with single-SNP analyses.

4.3 Reproducibility of DArTag Bulk Genotyping Across Pool Types

Genotyping of germplasm accessions adds significant value to collections by enabling efficient identification of diversity and structuring of breeding programs. For inbreeding species, individual plant genotyping is straightforward. Whereas, for outcrossing species such as alfalfa, where every plant is genetically unique, genotyping population bulks is a practical alternative that captures population-level allele frequencies at lower cost. A critical question is whether replicate bulk samples yield consistent results. In this study, we evaluated the reproducibility of DArTag genotyping across two independent 10-plant pools and a 20-plant composite bulk per accession. Allele frequency correlations between the two independent pools exceeded $r^2 = 0.90$ for most accessions, demonstrating high genotyping consistency. We found a high correlation between replicate pooled samples from each accession, with correlations of allele frequency matrix between the two ten plant samples for individual accessions mostly above 90%. This observation is consistent with the findings of Zhao et al. (2024) which found that the read count correlation of individual plant samples was not significantly different between three individual samples and 7 plant pooled samples. Even though the results in this study show high reproducibility between three types of pooled samples, comparing them with annual medics in Zhao et al. (2024) study requires further investigation. This is due to the difference in genetic structure of annual medics which shows higher levels of inbreeding that leads to strong genetic similarity among individuals within an accession.

4.4 Limitations of Bulk Genotyping and Rare Allele Detection

A limitation of bulk genotyping is reduced power to detect rare alleles or rare microhaplotypes, which requires individual plant sequencing. One limitation of individual plant sequencing in alfalfa is the many plants are needed to cover the genotypes in population due to high out crossing rate that leads to high genotyping cost (Medina et al., 2024). Genotyping individual plants is not useful for a population, since every plant in the population is genetically unique. Thus, pools are needed to keep costs contained, even though this will result in fewer rare alleles being identified. In this study, pooling 10 plants per population proved effective based on the high correlation of allele frequencies for all but one sample between replications. The cause of low correlation in this study warrants further investigation. Based on our results, pooling 10 plants per accession results in reliable DArTag genotypes, making the platform useful for large-scale genetic diversity studies in alfalfa.

Bulk genotyping has been validated as an effective approach for genetic studies of outcrossing forage crops (Annicchiarico et al., 2016; Kölliker et al., 2001). Beyond cost efficiency, pooling reduces genotyping noise from paralogous amplification and rare allele artifacts that can inflate individual-plant genotype calls (Medina et al., 2024).

4.5 Population Structure and Geographic Origin: STRUCTURE Analysis

The classification of germplasm in this study could be promising for California and southern US breeding programs. Interestingly, Clusters 1 and 2, despite being the smallest in size, showed similar or slightly higher numbers of alleles and effective alleles compared with other clusters, possibly reflecting the higher genetic diversity historically associated with the Levant and Iranian centers of alfalfa origin.

The STRUCTURE results support the hypothesis that the primary center of alfalfa origin lies within the regions of present-day Turkey and Iran. At $K = 2$, the population was partitioned into two distinct genetic clusters reflecting both geographic distribution and historical utility.

The first cluster comprises accessions from the Levantine, Turkish and Iranian regions. This group likely represents the primary source of alfalfa genetic diversity (Barnes, 1977; Veronesi et al., 2010).

The second cluster includes accessions from the Arabian Peninsula, Africa, India, and South America. The genetic similarity across these distant regions suggests a dispersal pattern driven by the ancient trade routes. Specifically, coastal access and maritime trade likely facilitated the distribution of alfalfa to regions bordering the Indian and Atlantic Oceans, allowing travelers and merchants to introduce the crop to the new continents.

STRUCTURE is a well-established tool for characterizing genetic backgrounds in complex organisms. In alfalfa, several studies have demonstrated the effectiveness of the STRUCTURE admixture model for investigating the genetic architecture of *Medicago falcata* and *M. sativa* accessions, including diploid wild germplasm (Şakiroğlu et al., 2010), tetraploid NPGS accessions spanning the *sativa-falcata* complex (İlhan et al., 2016), worldwide cultivated collections (Qiang et al., 2015), and populations under long-term selection (Wang et al., 2023). By delineating population boundaries and quantifying admixture proportions, STRUCTURE provides actionable information for breeders seeking to combine diverse germplasm sources, understanding critical for modern alfalfa breeding programs aiming to improve crop resilience and yield.

4.6 DArTag Microhaplotypes and PCA Corroboration

Population structure analyses in alfalfa have historically relied on SSR markers, which are labor-intensive and typically constrain sample sizes to fewer than 100 accessions (Annicchiarico et al., 2017; Qiang et al., 2015; Flajoulot et al., 2005; Kölliker et al., 2001). SNP markers address throughput limitations but their biallelic nature reduces per-locus polymorphism. The DArTag microhaplotype approach bridges these limitations: 2,769 of 3,000 loci passed depth and quality filters, and while one-third of target SNPs were monomorphic, only 4% of microhaplotype loci were monomorphic across the 376 accessions. High-throughput DArTag genotyping enabled population-scale analyses that would be impractical with SSRs.

The PCA results support the STRUCTURE findings. Accessions from the Levant, Arabian Peninsula, Africa, and India (STRUCTURE Clusters 1, 3, and 4) overlapped substantially in PCA space, consistent with extensive historical gene flow between these groups (Russelle, 2001). In contrast, there is a marked genetic distance from Turkey/Iran (cluster 2) and South America (cluster 5). These two clusters show limited gene flow due to the lack of diversity in these two clusters.

4.7 DAPC: Separation of South American and Iranian Accessions

DAPC further refined these relationships by maximizing discrimination among groups, clearly separating South American and Iranian accessions from the broader Arabian Peninsula, Africa, Levant, and India cluster. The genetic relatedness among Arabian Peninsula, African, and Indian accessions reflects centuries of maritime commerce across the Red Sea and Indian Ocean (Russelle, 2001). The distinct genetic background of South American accessions is consistent with a founder effect from Spanish introductions, while the unique diversity of Iranian accessions reflects localized adaptation during the Persian Empire period (Barnes, 1977; Ball, 2016).

4.8 Neighbor-Joining Tree and GRIN Passport Data Reliability

The Neighbor-Joining tree analysis provided a high-resolution view of phylogenetic relationships among individual accessions, broadly corroborating the cluster structure identified by STRUCTURE, PCA, and DAPC. However, a subset of accessions did not align with their expected geographic groups. This scattering of accessions suggests that the original passport data on GRIN (Germplasm Resources Information Network) may not always reflect the true origin of every accession. Two accessions from the Levant geographic group, for example PI 464765 and PI 464769, clustered within Cluster 2 (Iran and Turkey) rather than with other Levant accessions. In GRIN, both are listed under the common name "Iranian alfalfa," suggesting they may have originated from Iranian germplasm despite their Levant geographic assignment.

Discrepancies between passport data and genetic placement are not uncommon in genebank collections and can arise from several sources, including seed regeneration at non-native locations, mislabeling during collection or processing, accession misclassification, and the historical movement of germplasm along trade routes that predates formal collection records (Tanksley & McCouch, 1997). The case of PI 464765 and PI 464769 illustrates this well: their common name in GRIN directly contradicts their geographic assignment, pointing to a cataloguing inconsistency rather than a genuine genetic anomaly.

These findings underscore the importance of molecular genotyping as a complement to passport data when assembling diverse germplasm panels for breeding. Breeders who rely solely on geographic origin to select parents risk unknowingly including genetically misassigned accessions, which could reduce effective diversity in crossing programs. Integrating DArTag genotype data with passport records, as demonstrated here, provides a more reliable basis for germplasm characterization and strategic cross design in nondormant alfalfa improvement.

4.9 Intrapopulation Genetic Diversity and Heterozygosity Indices

Microhaplotype loci consistently showed higher intrapopulation genetic diversity than target SNPs, in agreement with Medina et al. (2024). The benefit was most pronounced in Cluster 2, where microhaplotypes reduced the proportion of monomorphic markers from 33.73% to 14.73%, a nearly 2.3-fold improvement. The alleles revealed exclusively through microhaplotype analysis represent diversity that would be entirely invisible to target-SNP-only approaches. This demonstrates the value of microhaplotypes for germplasm collections with narrow genetic bases, such as nondormant alfalfa, where extracting maximum information from available markers is essential for identifying useful variation.

Polyploids typically exhibit higher genetic diversity than diploids (Meirmans et al., 2018), and expected heterozygosity (H_s) was generally used to explain the genetic diversity of the population. (Nei, 1987)

In this study (H_s) was range from 0.25 to 0.27 for target SNPs and 0.29 to 0.31 for microhaplotypes. The expected heterozygous (H_s) results from this study were lower when compared to the genetic diversity study using SSR markers. Which the results were 0.62 - 0.72 for Qiang et al. (2015) 0.43 - 0.45 for Şakiroğlu et al. (2010). The lower (H_s) in DArTag study caused by the higher number of markers and loci which is different when compared to the SSR markers study which keep highly diverse loci only (Medina et al., 2024) (Zhao et al., 2024). However further investigation about the effect of pooled samples needs to be further investigated to accurately determine the genetic diversity indices.

The expected heterozygosity (H_s) for microhaplotypes is highest in cluster one (0.31) which include the most accessions from Levant and small portions of South America, Africa, Iran

and India countries while the other four clusters have equal H_s (0.29). The elevated diversity of Cluster 1 may reflect the historically high gene flow facilitated by the coastal geography of Levantine countries, which offered convenient access to maritime trade routes. The expected heterozygous range of pooled samples study (0.29 - 0.31) is consistent with the range of 0.21-0.31 reported by Annicchiarico et al. (2017) for pooled alfalfa samples, lending further support to the validity of the bulk genotyping approach.

4.10 The Analysis of Molecular Variance (AMOVA)

AMOVA partitioned total genetic variance into within-population (WP) and among-population (AP) components. The five STRUCTURE clusters explained 18% of total variance (AP), with 82% residing within populations (WP) for microhaplotypes. The AP component was notably higher than the 6% reported by Qiang et al. (2015) using SSR markers, though comparable to the 12.4% reported by Medina et al. (2024) using DArTag microhaplotypes. The predominance of within-population variance is typical of outcrossing species with high heterozygosity, frequent recombination, and strong inbreeding depression (Medina et al., 2024), and the consistently higher AP values in DArTag studies relative to SSR studies likely reflect the greater number of genome-wide loci captured by sequence-based platforms.

5. Conclusion

The DArTag platform proved an effective, high-throughput tool for characterizing the genetic diversity and population structure of large-scale germplasm collections of non-dormant alfalfa. This marker platform demonstrated the application of DArTag markers on the genetic diversity study of alfalfa and detection of unique allele from accessions in National Plant Germplasm collection by identifying the population structure and genetic diversity of 376 nondormant alfalfa accessions and checks into five clusters based on the geographic regions.

Although overall genetic diversity in nondormant alfalfa was relatively narrow compared with broader alfalfa germplasm panels, the analyses identified accessions carrying private and cluster-specific alleles that could enrich the genetic base of nondormant breeding populations. Microhaplotype markers consistently outperformed target SNPs in capturing this diversity, reducing the proportion of monomorphic markers from 33% to 4% and increasing total alleles detected. These results support the adoption of microhaplotype-based DArTag genotyping as a cost-efficient, information-rich approach for genetic diversity assessment in plant germplasm collections. The geographic and genetic framework established here provides alfalfa breeders in California and the southern US with a practical guide for selecting diverse parental germplasm, accelerating identification of useful alleles, and improving nondormant alfalfa cultivars.

6. Acknowledgement

This research was funded by grants from the USDA-Alfalfa Seed and Alfalfa Forage Research Program #Award Number 2019-70005-30361 to ECB and BI, from the California Crop Improvement Association to ECB, by USDA-ARS. Kreingkrai Nonkum is funded by a Royal Thai Government Scholarship. This project is part of Hatch Multistate Research Project NE-2210 (formerly NE1710).

7. References

- Annicchiarico, P., Wei, Y., & Brummer, E. C. (2017). Genetic structure of putative heterotic populations of alfalfa. *Plant breeding*, 136(5), 671–678.
<https://doi.org/10.1111/pbr.12511>
- Annicchiarico, P., Nazzicari, N., Ananta, A., Carelli, M., Wei, Y., & Brummer, E. C. (2016). Assessment of cultivar distinctness in alfalfa: A comparison of genotyping-by-sequencing, simple-sequence repeat marker, and morphophysiological observations. *The Plant Genome*, 9(2), 1–12. <https://doi.org/10.3835/plantgenome2015.10.0105>
- Ball, D. (2000). Little known facts about alfalfa.
https://uknowledge.uky.edu/ky_alfalfa/2000/Session/8/
- Barnes, D. K. 1977. *Alfalfa germplasm in the United States: Genetic vulnerability, use, improvement, and maintenance*. No. 1571. Department of Agriculture, Agricultural Research Service,

- Bingham, E. T., Groose, R. W., Woodfield, D. R., & Kidwell, K. K. (1994). Complementary gene interactions in alfalfa are greater in autotetraploids than diploids. *Crop Science*, 34(4), 823–829. <https://doi.org/10.2135/cropsci1994.0011183X003400030043x>
- Brummer, E. C., & Putnam, D. (2018). The alfalfa yield gap: What's holding us back? *Hay & Forage Grower* <https://www.hayandforage.com/article-1856-the-alfalfa-yield-gap-what%E2%80%99s-holding-us-back-.html> (2018).
- Brummer, E. Charles, M. Maroof Shah, and Diane Luth. (2000) "Reexamining the relationship between fall dormancy and winter hardiness in alfalfa." *Crop Science* 40, no. 4: 971-977.
- Brummer, E. Charles. (1999). "Capturing heterosis in forage crop cultivar development." *Crop science* 39, no. 4: 943-954.
- Brummer, E. C. (2002). Agronomic consequences of dormant-nondormant alfalfa mixtures. *Agronomy Journal*, 94(4), 782–785.
<https://doi.org/10.2134/agronj2002.0782><http://doi.org/10.2134/agronj2002.0782>
- Byrne, P. F., Volk, G. M., Gardner, C., Gore, M. A., Simon, P. W., Smith, S. (2018). Sustaining the future of plant breeding: The critical role of the USDA-ARS National Plant Germplasm System. *Crop Sci.* 58, 451–468. doi: 10.2135/cropsci2017.05.0303
<https://acsess.onlinelibrary.wiley.com/doi/full/10.2135/cropsci2017.05.0303>
- Chen, H.T. , Zeng, Y. , Yang, Y.Z. , Huang, L.L. , Tang, B.L. , Zhang, H. , Hao, F. *et al.* (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* 11, 2494.

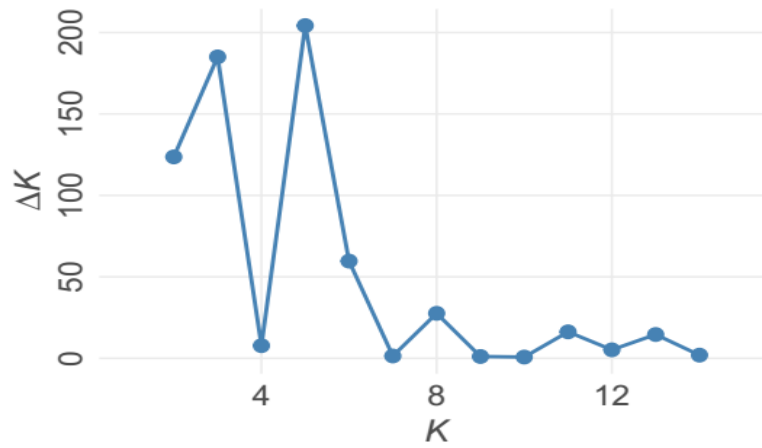
- Clark LV, Lipka AE, and Sacks EJ (2019). polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3* 9(3), 663–673.
<https://doi.org/10.1534/g3.118.200913>
- Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet. (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." *Molecular ecology* 14, no. 8: 2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
- Flajoulot, S., Ronfort, J., Baudouin, P., Barre, P., Huguet, T., Huyghe, C., & Julier, B. (2005). Genetic diversity among alfalfa (*Medicago sativa*) cultivars coming from a breeding program, using SSR markers. *Theoretical and applied genetics*, 111, 1420–1429.
<https://doi.org/10.1007/s00122-005-0074-4>
- Francis, R. M. (2017). "pophelper: an R package and web app to analyse and visualize population structure." *Molecular ecology resources* 17, no. 1 (2017): 27-32.
- Govindaraj M, Vetriventhan M, Srinivasan M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet Res Int*. 2015;2015:431487. doi: 10.1155/2015/431487. Epub 2015 Mar 19. PMID: 25874132; PMCID: PMC4383386.
- İlhan, D., Li, X., Brummer, E. C., & Şakiroğlu, M. (2016). Genetic diversity and population structure of tetraploid accessions of the *Medicago sativa–falcata* complex. *Crop Science*, 56(3), 1020–1029. <https://doi.org/10.2135/cropsci2015.12.0750>
- Irish, B.M., Greene, S.L. (2021). Germplasm Collection, Genetic Resources, and Gene Pools in Alfalfa. In: Yu, LX., Kole, C. (eds) *The Alfalfa Genome* . Compendium of Plant Genomes. Springer, Cham. https://doi.org/10.1007/978-3-030-74466-3_4

- Jombart, T., Devillard, S. & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 11, 94. <https://doi.org/10.1186/1471-2156-11-94>
- Köllicker, R., Jones, E. S., Jahufer, M. Z. Z., & Forster, J. W. (2001). Bulkcd AFLP analysis for the assessment of genetic diversity in white clover (*Trifolium repens* L.). *Euphytica*, *121*(3), 305–315. <https://doi.org/10.1023/A:1013368402655>
- Medina, Cesar A., Dongyan Zhao, Meng Lin, Manoj Sapkota, Alexander M. Sandercock, Craig T. Beil, Moira J. Sheehan et al. (2025). "Pre-breeding in alfalfa germplasm develops highly differentiated populations, as revealed by genome-wide microhaplotype markers. *Scientific Reports*, 15(1), 1253. <https://doi.org/10.1038/s41598-024-84262-x>
- Medina, C.A., Heuschele, D.J., Zhao, D., Lin, M., Beil, C.T., Sheehan, M.J. and Xu, Z., (2024). Multi-trait modeling and machine learning discover new markers associated with stem traits in alfalfa. *Frontiers in Plant Science*, *15*, p.1429976.
- Meirmans, P. G. (2020). GENODIVE version 3.0: easy-to-use software for the analysis of genetic data of diploids and polyploids. *Molecular Ecology Resources* *20*, 1126–1131
- Meirmans, P. G., & Liu, S. (2018). Analysis of Molecular Variance (AMOVA) for autopolyploids. *Frontiers in Ecology and Evolution*, *6*, 1–10. <https://doi.org/10.3389/fevo.2018.00042>
- NASS, USDA (2025). USDA quickstats. In: USDA National Agricultural Statistics Service. <https://quickstats.nass.usda.gov/> Accessed 20 July 2025

- Nei, M. (1987). *Molecular Evolutionary Genetics*. (Columbia University Press, New York, 1987). <https://doi.org/10.7312/nei-92038>
- Paradis, E., Claude, J., Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. <https://doi.org/10.1093/bioinformatics/btg412>
- Paradis E, Schliep K (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.” *Bioinformatics*, **35**, 526-528.[doi:10.1093/bioinformatics/bty633](https://doi.org/10.1093/bioinformatics/bty633).
- Pritchard JK, Stephens M, Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics*. Jun;155(2):945-59. <https://doi.org/10.1093/genetics/155.2.945>
- Postman, J., Hummer, K., Stover, E., Krueger, R., Forsline, P., Grau, L. J., et al. (2006). Fruit and nut genebanks in the US National Plant Germplasm System. *HortScience* 41, 1188–1194. <https://www.academia.edu/download/102826888/article-p1188.pdf>.
- Putnam, D., Meccage, E. (2022). Profitable alfalfa production sustains the environment in *Proceeding 2022 World alfalfa congress* (Davis, CA: University of California Davis), 14–17. : <https://alfalfasyposium.ucdavis.edu/+symposium/proceedings/2022/Articles/Daniel-Putnam-2022-Profitable-Alfalfa-Production-Sustains-The-Environment.pdf>
- Qiang, H., Chen, Z., Zhang, Z., Wang, X., Gao, H., Wang, Z. (2015). Molecular diversity and population structure of a worldwide collection of cultivated tetraploid alfalfa (*Medicago sativa* subsp. *sativa* L.) germplasm as revealed by microsatellite markers. *PloS One* 10, 1–12. <https://doi.org/10.1371/journal.pone.0124592>

- Russelle, M. P. (2001). Alfalfa: After an 8,000-year journey, the ‘Queen of Forages’ stands poised to enjoy renewed popularity. *American Scientist*, 89(3), 252–261.
<https://doi.org/10.1511/2001.3.252>
- Şakiroğlu, M., Doyle, J.J. and Charles Brummer, E., (2010). Inferring population structure and genetic diversity of broad range of wild diploid alfalfa (*Medicago sativa* L.) accessions using SSR markers. *Theoretical and applied genetics*, 121, pp. 403–415.
<https://doi.org/10.1007/s00122-010-1312-2>
- Şakiroğlu, M. and Brummer, E.C. (2007), Little Heterosis between Alfalfa Populations Derived from the Midwestern and Southwestern United States. *Crop Sci.*, 47: 2364-2371.
<https://doi.org/10.2135/cropsci2006.12.0803> .
- Segovia-Lerma A, Murray LW, Townsend MS, Ray IM. (2004). Population-based diallel analyses among nine historically recognized alfalfa germplasms. *Theor Appl Genet.* 2004 Nov; 109(8), 1568–1575. <https://doi.org/10.1007/s00122-004-1784-8> Epub 2004 Sep 15. PMID: 15372154.
- Simmonds, N. W. (1993). Introgression and incorporation. Strategies for the use of crop genetic resources. *Biological Reviews of the Cambridge Philosophical Society* 68, 539–562:
<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.1993.tb01243.x>
- Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G, Wu S, Wang Y. (2023). SRplot: A free online platform for data visualization and graphing. *PLoS One.* Nov 9; 18(11), e0294236. <https://doi.org/10.1371/journal.pone.0294236> PMID: 37943830.

- Tanksley SD, & McCouch SR. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science*. 1997 Aug 22;277(5329):1063-6. doi: 10.1126/science.277.5329.1063. PMID: 9262467.
- Teuber, L.R., K.L. Taggard, L.K. Gibbs, M.H. McCaslin, M.A. Peterson, and D.K. Barnes. (1998). Fall dormancy. *In* Standard tests to characterize alfalfa varieties. 3rd ed. (Amended 1998). North Am. Alfalfa Improvement Conf., Beltsville, MD. (Available online at <http://www.naaic.org/stdtests/Dormancy2.html>).
- Veronesi, F., Brummer, E. C., and Huyghe, C. (2010). Alfalfa. *In Fodder crops and amenity grasses* (pp. 395-437). New York, NY: Springer New York.
- Wang, Z., & Şakiroğlu, M. (2021). The Origin, Evolution, and Genetic Diversity of Alfalfa. *In*: Yu, LX., Kole, C. (eds) *The Alfalfa Genome . Compendium of Plant Genomes*. Springer, Cham. https://doi.org/10.1007/978-3-030-74466-3_3
- Wang H, Coulman B, Bai Y, Tar'an B, and Biligetu B. (2023). Genetic diversity and local adaption of alfalfa populations (*Medicago sativa* L.) under long-term grazing. *Sci Rep*. 2023 Jan 30;*13*(1), 1632. <https://doi.org/10.1038/s41598-023-28521-3> PMID: 36717619; PMCID: PMC9886962.
- Wilson, R. F., Stalker, H. T., & Brummer, E. C. (2004). Genomics research in alfalfa, *Medicago sativa* L. *Legume crop genomics*, 110-142.
- Zhao, Dongyan, Katherine Maria Mejia-Guerra, Marcelo Mollinari, Deborah Samac, Brian Irish, Katarzyna Heller-Uszynska, Craig Thomas Beil, and Moira Jane Sheehan. (2023) "A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.)." *In Genetic resources*, vol. 4, no. 8, pp. 55–63. <https://doi.org/10.46265/genresj.AQST8613>



Supplementary Figure 1. The determining of optimum value of K, the second order statistics (DeltaK) developed by Evanno et al. (2005)

Chapter 3

Identifying the Genetic Basis of Four Agronomic Traits in a Non-Dormant Alfalfa Germplasm Collection

Kreingkrui Nonkum, Matt Francis, Brian M. Irish, Scott Newell, Daniel Pap, Craig T. Beil,
Moir Sheehan, Dongyan Zhao, Meng Lin, and E. Charles Brummer

Abstract

The assessment of germplasm is the crucial step for plant breeding. In this study we use genotypic data combined with multi environment trials to evaluate 383 non-dormant alfalfa germplasm and varieties. Phenotypic data of dry weight, plant height, persistence and regrowth score were collected from three locations: Davis, CA, Five Points, CA, and Prosser, Washington, from 2021 to 2022. Single-trial and combined analysis were used to analyze phenotypic data and obtain BLUPs by environment. Genotyping was conducted using an alfalfa DArTag platform, resulting in a genotypic matrix comprising 2,769 single nucleotide polymorphisms (SNPs). A genome-wide association study found 69 markers for the four chosen agronomic traits from all models. *ArnA* like protein was found associated with markers for plant height and persistence. GBLUP method was used to obtain genomic prediction accuracies. The use of combined analysis using the BLUP and DArTag platforms showed low predictive ability (0.26) for dry weight while using single harvest BLUE show predictive ability of 0.17 to 0.75. This study shows the two approaches to improving alfalfa genomic study that could be useful for alfalfa yield improvement and germplasm assessments.

1. Introduction

Alfalfa (*Medicago sativa* L.) is known as “the queen of forages” due to the high productivity, excellent quality for animal feed, and nitrogen fixation. In the United States, the alfalfa estimated direct value in 2025 was \$7.6 billion and ranked first among forage crops by area, with a total of 14.6 million acres in 2025 (NASS, 2025). Cultivated alfalfa is an insect pollinated allogamous species; is autotetraploid ($2n=4x=32$), exhibiting tetrasomic inheritance; and has a haploid genome size of 816Mb (Chen et al., 2020, Shen et al., 2020, Long et al., 2022) Due to the allogamous nature of alfalfa, populations consist of many different genotypes expressing a high level of heterozygosity (Brummer and Casler, 2004). High heterozygosity probably enables alfalfa to adapt to various environments and survive multiple abiotic and biotic stresses. In Mediterranean environments such as California, alfalfa faces numerous threats, including limited water for irrigation, increased salinity, and higher temperatures (Lesk et al., 2016).

Fall dormancy refers to the response of alfalfa to decreasing photoperiod and temperature in autumn (Castonguay et al., 2006). Alfalfa cultivars are classified into dormancy classes based on the amount of regrowth produced in the autumn following a late-season harvest. A fall dormancy (FD) score of one represents very fall dormant and $FD = 11$ represents extremely non-dormant germplasm (Teuber et al., 1998). Alfalfa cultivars adapted to northern California, the upper midwestern USA and the northeastern USA are dormant to moderately dormant, with $FD = 2$ to 5 These are regions with cold winters, and thus cultivars require freezing tolerance. In contrast, cultivars adapted to the California Central Valley and the desert Southwestern USA are non-dormant, ranging in fall dormancy from $FD = 7$ to 11 . Fall dormant alfalfa has high winter hardiness but produces low yield during winter in regions with temperatures conducive to growth; nondormant alfalfa generally has

poor winter survival but grows throughout the year if conditions are favorable (Brummer, 2000; Brummer, 2002; Weishaar et al., 2005).

Plant gene banks contain alfalfa accessions that could be used to address challenges facing alfalfa production in the future. The reduced genetic diversity that occurs over time in a breeding program can decrease genetic gain and lead to slower adaptation to the environment, especially under the rapid climate change than if the program had greater diversity (Govindaraj et al., 2015). The genetic diversity from gene banks can be used via introgression or incorporation to increase diversity and bring new traits into breeding programs (Simmonds, 1993).

The National Plant Germplasm System of the U.S. Department of Agriculture (USDA), Agricultural Research Service (ARS), maintains plant germplasm resources of most crops, which can be searched using the GRIN-Global database (<https://www.grin-global.org/>) (Byrne et al., 2018; Postman et al., 2006). For alfalfa, the collection contains ~4,000 accessions of cultivated and wild relatives of alfalfa (Teuber et al., 1998). Due to the large number of accessions in the system, effectively using them in alfalfa improvement is difficult. Pre-breeding, or developing new breeding populations from germplasm collections, is one way to access the genetic diversity present in the collection to improve quantitative traits (Simmonds, 1993).

Alfalfa breeding uses phenotypic recurrent selection for population improvement by evaluating genetic breeding value of plants under field conditions. The advancement of high throughput genotyping allows alfalfa breeders to apply genomic prediction to alfalfa selection programs and Genome-Wide Association Study (GWAS) to help identify quantitative trait loci that can be selected by marker-assisted selection.

Although past studies, such as selection mapping in alfalfa (Munjal et al., 2018), used the *M. truncatula* genome as a reference, genome sequencing of cultivated *M. sativa*, ‘XinJiangDaYe’ (Chen et al., 2020), ‘Zhongmu 1’ (Shen et al., 2020) and ‘Zhongmu 4’ (Long et al., 2022) now enables researchers to use the alfalfa genome as a reference for GWAS and genomic prediction experiments that use genotyping-by-sequencing (GBS) (Elshire et al., 2011) or other sequence-based marker platforms. Recent whole genome marker studies have applied GWAS and genomic prediction to alfalfa grown under drought stress (Medina et al., 2025) and for stem traits (Medina et al., 2024) and evaluated the genetic diversity of alfalfa germplasm (Pégard et al., 2023). A Diversity Array Technology genotyping platform (DarTag) for alfalfa was recently developed by the USDA-ARS, Breeding Insight program (Zhao et al., 2023). The DarTag platform identifies the sequences surrounding 3,000 SNP loci distributed across alfalfa genome. These flanking sequences of 81 base pairs in the current version of the method can be used to identify additional off-target SNP, the entire sequence is referred to as a microhaplotype (Medina et al., 2024), which provides multi-allelic information in the vicinity of each target SNP locus. Many haplotypes can be found per locus.

Multi-environment-Trials are essential for estimating genotype performance across varied conditions and determining the significance of genotype by environment interaction ($G \times E$) (Malosetti et al., 2013). Mixed Linear Models are often used for multi-environment-trials analysis (Smith et al., 2015) due to their flexibility to include random effects in the model. The multi-environment trials can be analyzed by two approaches: single stage and stagewise analysis (Lin et al., 2023). A single stage model includes all sources of variation into one mixed linear model to obtain best unbiased linear estimates (BLUEs) or best linear unbiased predictions (BLUPs) (Smith et al., 2001). However, single stage analysis requires complex models due to the specific details

of spatial design and may need high computational power. To overcome the single stage issue, the stagewise analysis can be used by separating the analysis into two stages to reduce model complexity and computational burden (Möhring and Piepho 2009; Damesa et al. 2017). In a stagewise analysis, spatial adjustment is first applied within the first stage to account for field heterogeneity, and spatially adjusted BLUEs are then obtained as output. These spatially adjusted BLUEs, along with their associated variance and covariance matrices, are subsequently used as inputs in the second stage, which incorporates repeated measures across harvests and locations to obtain BLUPs representing the final adjusted genotypic means (Möhring and Piepho 2009).

Given that non-dormant germplasm represents an important reservoir of genetic diversity for improving alfalfa in non-dormant production systems where heat tolerance, salt tolerance, and high-yield potential under irrigation are priority traits, the evaluation of this resource is both timely and necessary. We evaluated a collection of 383 non-dormant alfalfa accessions that were previously genotyped to assess their population structure (Nonkum et al., Chapter 2). These accessions were planted in field trials at Davis, CA; Five Points, CA; and Prosser, WA. The aims of this study were: (i) to evaluate phenotypic variation among non-dormant alfalfa accessions for dry weight, plant height, persistence, and regrowth across multiple environments and harvest dates using a stagewise (multi-environment trials approach); (ii) to identify SNP markers associated with these four agronomic traits using a genome-wide association study (GWAS); and (iii) to assess the predictive ability of the 3K DArTag marker platform for genomic prediction of these traits.

2. Materials and Methods

2.1 Plant Materials and Field Trials

A set of 383 non-dormant alfalfa accessions from regions with mild winter climates, including the Arabian Peninsula, India, Iran, the Levant (plus southern Turkey) region, North Africa, and South America (Supplementary File 1), were obtained from the USDA–National Plant Germplasm System (NPGS). This germplasm represents a promising genetic resource for improving alfalfa adapted to western and southern US production systems. In addition, fall dormancy check cultivars (Teuber et al., 1988) from FD=4-11, and three check entries (WL 365HQ', AmeriStand 803T', and UF_AlfPers_2015 (an experimental from the University of Florida)) were included for comparison. Seeds of all accessions and check entries were sown in the greenhouse and transplanted to field trials in spring 2021 at the UC Davis Plant Sciences Farm in Davis, CA (38°31'39.2" N and 121°46'21.0" W), the UC Westside Research and Extension Center in Five Points, CA (36°20'15.0" N and 120°6'37.4" W) and the Washington State University Irrigated Agriculture Research and Extension Center in Prosser, WA (46°15'24.6" N and 119°43'50.8" W).

Each field trial was established using a randomized complete block design with two replicates. Each plot contained five plants with 30 cm between individual plants and 75 cm between plots. Border rows were planted with the same spacing surrounding the entire plot area. Each trial consisted of 800 plots because the three check entries were included three times in each replication. These were arranged in a 20×40 grid (columns by rows) at Davis, and a 40×20 grid at Five Points and Prosser. The trials were managed under full irrigation at all three locations. The irrigation water in Five Points was saline ($EC \sim 8.5 \text{ dS m}^{-1}$) for evaluating performance under salt

stress. Prosser had a cold winter, useful for evaluating nondormant alfalfa for winter survival. Alfalfa was harvested by hand from each trial plot at approximately 30-35 day intervals throughout the growing season.

2.2 Phenotypic data collection

Data were collected at specified points during the experiment, on selected harvests (Table 1). Four traits including plant height, persistence, dry weight, and regrowth were measured during 2021 and 2022. One week before harvest, the plant height of each plant in each plot was measured based on the natural standing height and the average height per plot was computed. Forage biomass per plot was measured by hand harvesting aboveground plant material approximately 5 cm above the soil surface for the entire plot, placing it in a bag, and drying it for in a hot air oven at 60°C for one week and dry weight per plot was adjusted to dry weight. Regrowth was scored one week after harvest on a scale of 1 = extremely slow regrowth and 5 = extremely fast and uniform regrowth. Persistence was computed by determining the percentage of plants per plot surviving at the last stand count compared to the number at the first stand count, thereby excluding any plants lost during the establishment phase.

Table 1. Dates on which four traits were measured in Davis, CA; Five Points (Westside), CA; and Prosser, WA.

Location	Harvest date	Season	Trait			
			Dry Weight	Plant Height	Persistence	Regrowth
Davis1	Sep, 2021	Fall	Yes	Yes	NA	NA
Davis2	Oct, 2021	Fall	Yes	Yes	Yes	Yes
Davis3	April, 2022	Spring	Yes	Yes	Yes	Yes
Davis4	May, 2022	Spring	Yes	Yes	Yes	Yes
Davis5	June, 2022	Spring	Yes	NA	Yes	NA
Prosser1	Oct, 2021	Fall	Yes	Yes	Yes	Yes
Prosser2	April, 2022	Spring	Yes	NA	Yes	NA
Prosser3	Oct, 2022	Fall	Yes	Yes	Yes	Yes
Westside1	July, 2022	Summer	Yes	NA	Yes	NA
Westside2	Oct, 2022	Fall	Yes	Yes	Yes	Yes

2.3 Phenotypic data analysis

Phenotypic traits within each location were analyzed at each collection time point using a mixed linear model with the SpATS package v. 1.0-18 (Rodríguez et al., 2018) to account for spatial variation across the trial. This model adjusts the phenotypic data with two-dimensional penalized splines to obtain best linear unbiased estimates (BLUEs):

$$y = X\beta + f(r,c) + Z_u u + Z_g g + \varepsilon$$

where y represents the phenotypic response of each of the 800 plots, X is the association design matrix and β is a vector of fixed effects. The function $f(r,c)$ represents a bivariate function of rows and columns corresponding to the vector of spatial effects. u is a vector of random row and column effects with Z_u is the associated incidence matrix. g is a genotypic vector of 383 accessions, with Z_g as the associated matrix assign as fixed effect. ε is an error random vector of each plot, which is distributed $\varepsilon \sim N(0, \sigma^2 \varepsilon I_{800})$. Broad-sense heritability was calculated using the function `getHeritability` from the SpATS R package v. 1.0-18 (Rodríguez et al., 2018)

To assess the statistical significance of genotypic effects within each environment, a variance component Z-test was conducted using the output of the SpATS random genotypic model. The genotypic variance component (σ_g^2) was extracted directly from the fitted model for each trait–environment combination and its standard error was approximated by combining the asymptotic REML variance form for variance components (Searle et al., 1992, Chapter 6) and a PEV-based correction for BLUP shrinkage, following the identity $\text{Var}(w) + \text{mean}(\text{PEV}) = \sigma_g^2$

(Mrode, 2005), where PEV quantifies estimation uncertainty in predicted genotypic effects obtained from the diagonal of C_{11}^{-1} (Gilmour et al., 1995):

$$SE(\sigma_g^2) = \sqrt{\frac{2}{ED_g^2}(\sigma_g^2 + \text{mean}(PEV)^2)}$$

where ED_g^2 is the effective dimension of the genotype term from the SpATS model, reflecting the number of independently estimated genotypic effects after spatial regularization, and $\text{mean}(PEV)^2$ is the average prediction error variance across all genotypes, obtained from the diagonal of the generalized inverse of the genotype coefficient matrix (C_{11}^{-1}) stored in the SpATS model object. The Z-statistics were then computed as

$$Z = \frac{\sigma_g^2}{SE(\sigma_g^2)}$$

a one-sided p-value was derived as $P = 1 - \Phi(Z)$, where Φ denotes the standard normal cumulative distribution function. The one-sided test is appropriate because σ_g^2 is bounded at zero by definition and the hypothesis of interest is $H_0: \sigma_g^2 = 0$ against $H_1: \sigma_g^2 > 0$.

We used a stagewise model (ST2), for analysis across data collection dates (considering all harvest/environment combinations as independent) by taking single-environment BLUEs as input to generate best linear unbiased predictors (BLUPs) using lme4breeding in R (Covarrubias, 2025). The stagewise approach enabled estimation of genotypic, environment, and genotype-by-environment interaction effects in a nested model as follows:

$$y_{ij} = \mu + E_j + G(E)_{ij} + \varepsilon_{ij}$$

where y_{ij} is the response variable observed in i th accession at the j th environment; μ is the overall mean; E_j is the fixed effect of j th environment; $G(E)_{ij}$ is a vector of random genetic effects of accession i nested in the j th environment; assuming a distribution of $G(E)_{ij}$ identical and independently distributed, $G(E)_{ij} \sim N(0, \sigma_{G(E)}^2)$ with variance covariance matrix assumption of unequal variance and/or covariances between environments.

Using lme4breeding, we compared a model including the genotype main effect (GENO) with a model containing the genotype main effect and genotype $\times$ environment interaction (G $\times$ E). using the Akaike Information Criterion (AIC). The GENO model assumed that the relative performance of genotypes remained consistent across environments while the G $\times$ E model accounted for the interaction between genotypic and environmental effects. We use the Akaike Information Criterion (AIC) instead of the Bayesian Information Criterion (BIC) because BIC imposes a stronger penalty for model complexity that scales with sample size and can systematically favor simpler models even when biologically meaningful genotype-by-environment interaction is present (Burnham & Anderson, 2004).

2.4 DArTag genotyping

Genotyping using the DarTag SNP array was described in Chapter 2. Briefly, for each entry, two ten-plant samples were extracted and then pooled for genotyping. DNA was extracted at the USDA-ARS, Prosser WA using DNeasy 96 plant kits (Qiagen, 69181) samples were shipped to Diversity Array Technologies (DArT; www.diversityarrays.com, Australia) for DArTag genotyping using 3K platform developed by Breeding Insight, (Ithaca, NY, USA) of the USDA ARS (Zhao et al., 2023). DArTag is an amplicon-based targeted genotyping platform that sequences regions flanking the target SNP loci and can also identify off-target SNPs to generate microhaplotypes at each locus. The DArTag report named missing allele discovery counts (MADC) contains the reference and alternative allele read counts for target SNPs and of the multiple microhaplotype alleles. Further details on the genotyping are available in Chapter 2 of Nonkum et al. (2026).

2.5 Filtering and allele dosage estimation

The marker loci and associated microhaplotypes with a total read count ≥ 10 in $\geq 95\%$ of the samples were retained for genotype calling. The filtered MADC file with genomic regions and observed read counts was imported to R using the function ‘readDArTag’ in the PolyRAD v.2.0.0 R package (Clark et al., 2019). MADC file was used as an input file for genotype estimation by using ‘IteratePopStructure’ function in PolyRAD. The final genotype calls at the microhaplotype level were exported as a Variant Call Format (VCF) file using the ‘RADdata2VCF’ function with option `asSNPs = False`. Targeted SNP genotypes were called from the filtered MADC dataset by identifying both reference and alternative allele. Although the individual DNA samples represented bulks of genotypes, we called a single “genotype” for each sample corresponding to

reference allele frequencies. A SNP diversity plot was generated using the Cmpplot website-based application (Tang et al., 2023)

2.6 Population structure

Population structure analysis was performed as described in Chapter 2, section 2.6. The analysis was performed using Principal Component Analysis (PCA) by using GenoDive v.3.0 (Meirmans, 2020). Only target SNPs with a biallelic allele VCF file were imported to create RADobject using ‘VCF2RADdata’ function. The RADdata object was converted to the Structure (Clark et al., 2019) format file using ‘Export_Structure’ in PolyRAD.

2.7 Association Mapping Study

The association mapping analysis was performed using the R package GWASpoly (Rosyara et al., 2016) with a Q + K linear mixed model. The population structure (Q) was based on principal components and the correlations among genotypes were derived from the kinship matrix (K) as follows (Rosyara et al., 2016):

$$y = \mathbf{X}\beta + \mathbf{ZS}\tau + \mathbf{ZQ}v + \mathbf{Z}u + \varepsilon$$

where y corresponds to a vector of observed phenotypes; β is a vector of fixed effects; $\mathbf{X}$ is an incidence matrix used to model environmental effects; v is the subpopulation vector effects; $\mathbf{Q}$ is an incidence matrix for a population structure of size m ; u is a polygenic effect vector. $\mathbf{Z}$ is a matrix of incidence mapping genotypes to phenotypic observation; τ is a vector of SNP effect; $\mathbf{S}$ is a structure

incidence matrix; and ε is a residual vector. Markers associated with the four traits were identified using a Bonferroni-corrected significance threshold of $\alpha < 0.05$. Associated markers were annotated using alfalfa pan-transcriptome information (Medina et al., 2021). Genes in the same location of the SNP marker were annotated using the UniProt database (Bateman, 2019) using the reference transcriptome dataset for *M. sativa* (Medina et al., 2021)

2.8 Genomic selection

Phenotypic values of plant height and dry weight at the individual harvest level (ST0) and overall harvests (ST2) were used to estimate genomic estimated breeding values (GEBVs) based on genomic best linear unbiased prediction (GBLUP). GBLUP used the genomic relationship matrix (G matrix) (VanRaden, 2008) that was computed from genetic marker data as follows:

$$G = \frac{ZZ'}{2\sum p_i(1 - p_i)}$$

where Z is an incidence matrix for the DArTag markers and p_i is the observed minor allele frequency. The genomic relationship matrix of autotetraploid alfalfa was calculated using AGHmatrix in R (Amadeu et al., 2023) based on 2,769 DarTag. Genomic prediction (GP) models were trained for 100 iterations using 5-fold cross validation with 80% of genotypes assigned to the training dataset and the remaining 20% of genotypes being predicted as the testing dataset using lme4breeding in R (Covarrubias, 2025). The accuracy of the prediction model was calculated as the Pearson correlation coefficient between the predicted GEBVs of the testing set and their observed phenotypic values.

3. Results

3.1 SNP distribution

After applying a missing-data threshold of less than 5%, 2,769 out of the 3,000 marker loci were retained for downstream analysis, and the markers were distributed across all eight chromosomes of the alfalfa genome, although marker density varied within and among chromosomes (Nonkum, 2026, Chapter 2, section 3.1).

3.2 Phenotypic data summary

In this trial, we evaluated 383 germplasm accessions collected from alfalfa growing regions with mild winter climates. We did not know the fall dormancy ratings of all the germplasm, but we assumed most would have a dormancy of $FD \geq 6$. We assessed variation in biomass yield, plant height, persistence, and regrowth, including autumn regrowth, which is the determinant in assigning a fall dormancy class, across three locations with large differences in environment. Spatial variation potentially affects phenotypic variation, so we used the SpATS R package (Rodríguez et al., 2018) to model spatial variation at each time point in each location to adjust phenotypic data prior to further analysis.

Significant genotypic variation was detected for all four traits at every harvest time point across all three locations, based on variance component Z-values ($P < 0.001$) (Table 2). Z-values ranged from 16.0 (Persistence, Westside1) to 255.7 (Plant height, Davis1), with higher Z-values generally corresponding to higher broad-sense heritability estimates. Broad-sense heritability (H^2) varied by trait and environment (Table 2). Plant height showed consistently high heritability across all locations and harvests ($H^2 = 0.73$ – 0.96). Dry weight showed moderate heritability at most locations ($H^2 = 0.43$ – 0.73), with the exception of the two Westside harvests, where H^2 was markedly lower (0.14 – 0.43). Regrowth had high heritability at most timepoints ($H^2 = 0.65$ – 0.84) but was

notably low in fall harvests Davis2 ($H^2 = 0.14$) and Prosser3 ($H^2 = 0.14$). Persistence showed moderate heritability overall ($H^2 = 0.11$ – 0.78), with the lowest values at Westside ($H^2 = 0.11$ – 0.38).

Among accessions, phenotypic means and ranges were broad for all traits (Table 2). At Davis, dry weight ranged from 5 - 231 g (Davis1), and at Prosser, from 0.2 to 717 g in spring 2022 (Prosser2). Plant height ranged from 3.4 to 52.3 cm at the first Davis harvest and 6.7 to 64.1 cm at the first Prosser harvest. Regrowth scores spanned nearly the full scale at most environments, from near zero to above 5, while persistence ranged from as low as 5% to as high as 100% across Davis environments

Across locations, genotypic performance varied considerably, reflecting the contrasting production environments at Davis, Prosser, and Five Points. Dry weight was highest at Prosser2 (mean = 153 g plant⁻¹) and lowest at Five Points (Westside2) under saline irrigation (mean = 29 g plant⁻¹) (Table 2). Plant height means were substantially lower at Westside2 (mean = 10 cm) compared to Davis and Prosser environments (means of 15.5–65.5 cm). Regrowth was consistently lower in the cooler Prosser environment (means of 1.6–2.4) relative to Davis and Westside (means of 2.7–3.7). Persistence declined over time at all locations, with Prosser3 showing the greatest average stand loss (mean = 50%), likely associated with the effect of cold winter conditions on non-dormant accessions.

Table 2. Broad-sense heritability and significance test for the presence of variance among genotypes and adjusted phenotypic mean variation for four phenotypic traits measured in multiple environments.

Environment	H ²	Z-value	P-value	Mean	Range
Dry Weight (g plant⁻¹)					
Davis1	0.67	139.5	<0.001	70.44	5.07 – 230.73
Davis2	0.68	141.9	<0.001	41.93	0.83 – 90.86
Davis3	0.71	148.2	<0.001	30.32	0.42 – 103.49
Davis4	0.49	87.8	<0.001	42.34	8.77 – 111.78
Davis5	0.60	112.6	<0.001	86.83	6.77 – 209.22
Prosser1	0.65	132.2	<0.001	59.57	25.34 – 122.18
Prosser2	0.73	133.4	<0.001	152.77	0.16 – 716.73
Prosser3	0.69	122.7	<0.001	109.89	12.12 – 514.55
Westside1	0.43	77.1	<0.001	45.46	1.12 – 143.12
Westside2	0.14	20.8	<0.001	29.23	0.46 – 113.29
Plant Height (cm)					
Davis1	0.96	255.7	<0.001	25.91	3.48 – 52.3
Davis2	0.73	158.4	<0.001	64.64	14.33 – 96.84
Davis3	0.81	182.3	<0.001	65.52	11.12 – 104.93
Davis4	0.85	190.2	<0.001	15.51	1.6 – 41.71
Prosser1	0.95	250.3	<0.001	32.45	6.71 – 64.14
Prosser3	0.88	183.0	<0.001	22.23	4.22 – 58.81
Westside2	0.87	212.3	<0.001	10.37	1.85 – 28.97
Regrowth (score)					
Davis2	0.14	20.4	<0.001	2.67	-0.03 – 5.19
Davis3	0.70	148.6	<0.001	2.66	-0.04 – 6.46
Davis4	0.72	154.5	<0.001	3.24	-0.04 – 5.52
Prosser1	0.84	200.7	<0.001	2.36	0.73 – 6.05
Prosser3	0.14	17.1	<0.001	1.56	-0.14 – 5.68
Westside2	0.65	132.4	<0.001	3.68	0.68 – 6.09
Persistence (%)					
Davis1	0.54	102.4	<0.001	87.01	19.74 – 100.00
Davis2	0.30	47.7	<0.001	91.39	4.98 – 100.00
Davis3	0.44	78.1	<0.001	74.14	17 – 100.00
Davis4	0.46	81.5	<0.001	74.15	17 – 100.00
Davis5	0.52	93.9	<0.001	63.74	14.29 – 100.00
Prosser1	0.50	91.8	<0.001	85.77	27.13 – 100.00
Prosser2	0.78	146.7	<0.001	62.64	19.32 – 100.00
Prosser3	0.73	136.2	<0.001	49.56	16.89 – 100.00
Westside1	0.11	16.0	<0.001	72.94	26.33 – 100.00
Westside2	0.38	65.5	<0.001	70.00	17.76 – 100.00

H² = broad-sense heritability from SpATS random model. Z-value = $\sigma^2g / SE(\sigma^2g)$ from variance component Z-test (H₀: $\sigma^2g = 0$); P-value is one-sided. Mean and range on original (back-transformed) scale.

3.3 Multi-Environment Trial Analysis and Model Selection

To combine information across the ten harvest time points and three locations, we used a stagewise analysis approach in which the spatially adjusted BLUEs from each harvest were used as inputs to a second-stage mixed model. The second stage estimated BLUPs by accession across all environments (overall harvest). Prior to computing overall harvest BLUPs, the Akaike Information Criterion (AIC) was used to compare a genotype main-effect model (GENO) with a genotype-by-environment interaction model ($G \times E$) for each trait (Table 2). AIC was preferred over the Bayesian Information Criterion (BIC) because BIC imposes a stronger penalty for model complexity that increases with sample size (383 accessions in this experiment). In a multi-environment trials context where the primary goal is selecting the model that has relevant $G \times E$ structure rather than minimizing parameter count, AIC's more moderate penalty provides a better balance between goodness of fit and model complexity (Burnham & Anderson, 2004).

Table 3. Summary of model performance comparisons between the Genotype (GENO) and Genotype-by-Environment (G×E) models for four agronomic traits. Evaluated parameters include log-likelihood (logLike), AIC, Bayesian Information Criterion (BIC), and variance components for both GE and G×E effects. Lower AIC values served as the primary indicator for superior model fit.

Trait	Fixed Effects	Random Effects	logLike	AIC	BIC	Variance: G	Variance: G×E
Dry weight	Env	GE	-5109.35	10264.71	10421.91	3.45	-
	Env	G×E	-5520.18	11070.35	11172.87	6.37	0.08
Plant height	Env	GE	-434.53	903.06	1013.32	7.532	-
	Env	G×E	-829.96	1683.92	1761.75	18.563	0.046
Persistence	Env	GE	-2224.12	4490.24	4631.6	0.44	-
	Env	G×E	-2492.73	5013.46	5107.7	0.01	0.037
Regrowth	Env	GE	-1897.54	3831.07	3945.58	6.826	-
	Env	G×E	-2060.78	4143.56	4213.54	9.606	0.035

The GENO model showed a lower AIC than the G×E model for dry weight, persistence, and regrowth score (Table 3), indicating that genotype rankings for these traits were relatively consistent across environments and that a model without an explicit G×E term was more parsimonious. In contrast, the G×E model provided a marginally better fit for plant height, suggesting that the relative ranking of accessions for this trait differed across the three contrasting environments, a result consistent

with the environmental differences in mean plant height observed among Davis (30.3 cm), Prosser (40.1 cm), and Westside (20.5 cm). The G×E model was therefore used for plant height BLUPs in overall harvest model, while the GENO model was used for the remaining three traits.

3.4 Pearson Correlations Among Traits

Pearson correlations of overall harvest BLUPs from the GENO model revealed moderate positive relationships among dry weight, plant height, and regrowth score, consistent with the biological expectation that taller, faster-growing plants also tend to produce more biomass (Figure 1). In contrast, persistence was weakly or negatively correlated with dry weight and plant height. The strongest pairwise correlation was between plant height and regrowth score at Prosser ($r = 0.79$), while the weakest was between persistence and plant height at the same location ($r = -0.33$).

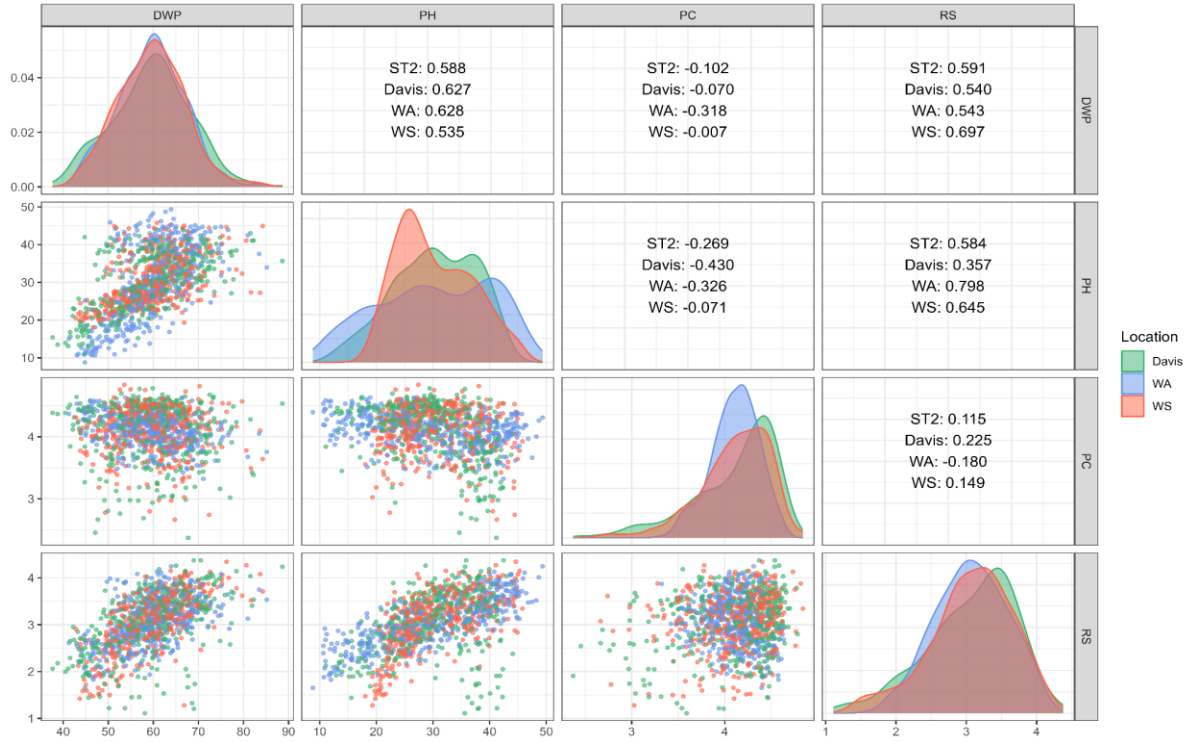


Figure 1. Pearson correlation matrix and scatter plots of overall BLUPs for dry weight (DWP), plant height (PH), persistence (PC), and plant regrowth (RS). The upper right diagonals display the Pearson correlation coefficients (r) calculated for each trait within and across environments (Davis, CA; Prosser, WA; Five Points (WS), CA). The lower left diagonal contains bivariate scatter plot illustrating the distribution of predicted BLUPs for all 383 accessions in three environments.

3.5 Genome-Wide Association Study Results

GWAS was conducted using spatially adjusted BLUEs from single harvest time points (ST0) and overall BLUPs as response variables, with the Q+K mixed model in GWASpoly. A total of 69 unique marker-trait associations (60 unique markers) were identified across all models and time points at a Bonferroni-corrected threshold of $\alpha < 0.05$ (Supplementary Table 1).

The largest number of associations were detected using single harvest BLUEs from individual harvests (50 marker-trait pairs), while overall harvest BLUPs identified 24 marker-trait pairs; 5 marker-trait pairs were detected in both. The 60 unique markers were distributed across all chromosomes: Chromosome 1 (11 markers), Chromosome 2 (8), Chromosome 3 (8), Chromosome 4 (8), Chromosome 5 (7), Chromosome 6 (5), Chromosome 7 (3), and Chromosome 8 (10).

Of the 60 unique markers, 31 were located in annotated coding regions and 29 in non-coding regions. The proportion of non-coding markers varied by trait: dry weight had 14 coding and 11 non-coding markers, plant height had 6 coding and 6 non-coding, persistence had 5 coding and 10 non-coding, and regrowth scored 6 coding and 8 non-coding markers. The overall harvest combined analysis used a GENO multi-environment model for dry weight, persistence, and regrowth score, but a genotype-by-environment interaction (GxE) model for plant height, reflecting greater genotype-by-environment sensitivity for plant height. This pattern suggests that harvest-specific environmental effects modulate marker detection, and that single-harvest analyses capture transient or environment-specific QTL that are averaged out in the combined overall harvest analysis. Five GWASpoly models, General, Diplo-general, additive, diploid-additive, 1-Dom and 2-Dom were tested in this study.

3.6 Markers Associated with Dry Weight

Twenty-five unique markers were associated with dry weight across harvests. The most stable marker was Chr6_1_080096025 (Chr6, position 80,096,025), annotated to a *TIR-NBS-LRR RCT1* resistance protein (Fragment), detected across four harvest environments Davis1, Davis2, Davis3, and overall harvest (ST2 GENO) and also significant for plant height. Among other coding-region markers, Chr8_084174416 was annotated to protein FATTY ACID EXPORT 2, chloroplastic-like, and Chr1_075444453 was annotated to PHD finger protein At1g33420-like. A QTL peak on Chromosome 6 was apparent in the Davis overall analysis Manhattan plot (Supplementary Table 1), where the additive model estimated an increase of 8.79–17.42 g plant⁻¹ associated with Chr6_1_080096025 across the Davis harvests (Davis overall: 8.79 g plant⁻¹, Davis1: 17.42 g plant⁻¹, Davis2: 12.61 g plant⁻¹, Davis3: 14.05 g plant⁻¹). The recurrence of Chr6_1_080096025 across four independent harvest environments (Figure 2), and its detection under all five genetic models tested (general, additive, diplo-additive, diplo-general, and 1-dom-alt), makes this locus a strong candidate for further investigation. A second marker, Chr8_1_014266142 (Chr8, position 14,266,142), was also detected across all five genetic models in Davis1 (Score = 6.31) and showed the largest single-harvest additive effect for dry weight in this study (+23.06 g plant⁻¹), making it another high-priority candidate despite its detection in only one harvest environment. Among dry weight markers with large effects, notable negative-effect loci were also identified: Chr5_1_020147516 (Chr5) showed a -38.1 g plant⁻¹ effect under the diplo-additive and 1-dom-alt models in the Prosser1 environment, and Chr3_1_075963488 (Chr3), annotated to protein POLLEN DEFECTIVE IN GUIDANCE 1 isoform X2, showed a -99.8 g plant⁻¹ effect under the 2-dom-alt model in the Prosser2 environment. These large negative effects

identify alleles that substantially reduce dry weight and should be considered in marker-assisted selection alongside positive-effect loci.

3.7 Markers Associated with Plant height

A total of 13 unique markers were associated with plant height, distributed across Chromosomes 1 (4 markers), 2 (2), 3 (1), 5 (2), 6 (3), and 7 (1). A prominent marker on Chromosome 8 (Chr8_011413748) was annotated to a *TIR-NBS-LRR* disease resistance protein and was also detected for dry weight. The same marker Chr4_044864682, annotated to an *ArnA*-like protein (Supplementary Table 1), was shared with dry weight, persistence, and regrowth, suggesting pleiotropic effects at this locus. The marker Chr6_1_080096025 (Chr6) was shared between dry weight and plant height, detected in Davis3 and Davis overall environments for plant height, further underscoring the cross-trait relevance of this Chromosome 6 locus (Figure 3). Additional coding-region plant height markers included Chr7_077351871, annotated to an NBS resistance protein (Fragment); Chr5_025333490, annotated to a TRAF-type domain-containing protein; and Chr6_050598088, annotated to an uncharacterized protein.

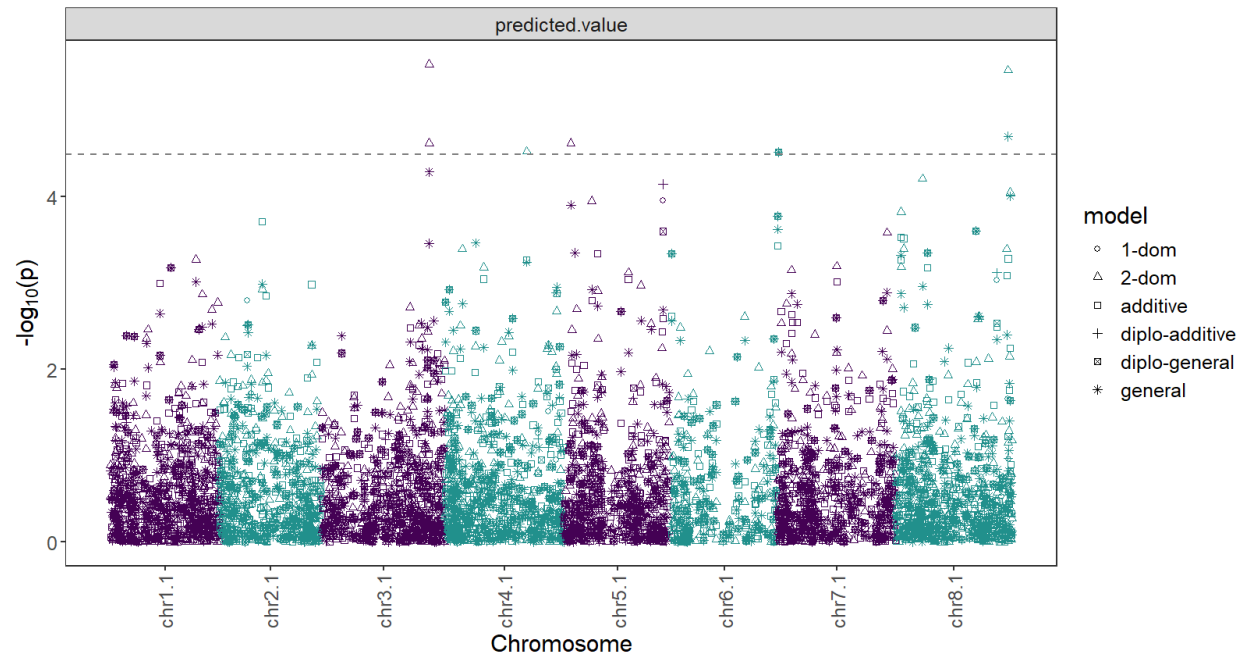


Figure 2. The Manhattan plot of GWAS model for dry weight using overall harvest (ST2) BLUPs across 2,769 DarTag marker, the marker Chr8_011413748 was associated with Disease resistance protein (Fragment) *NBS-LRR* gene for dry weight and plant height.

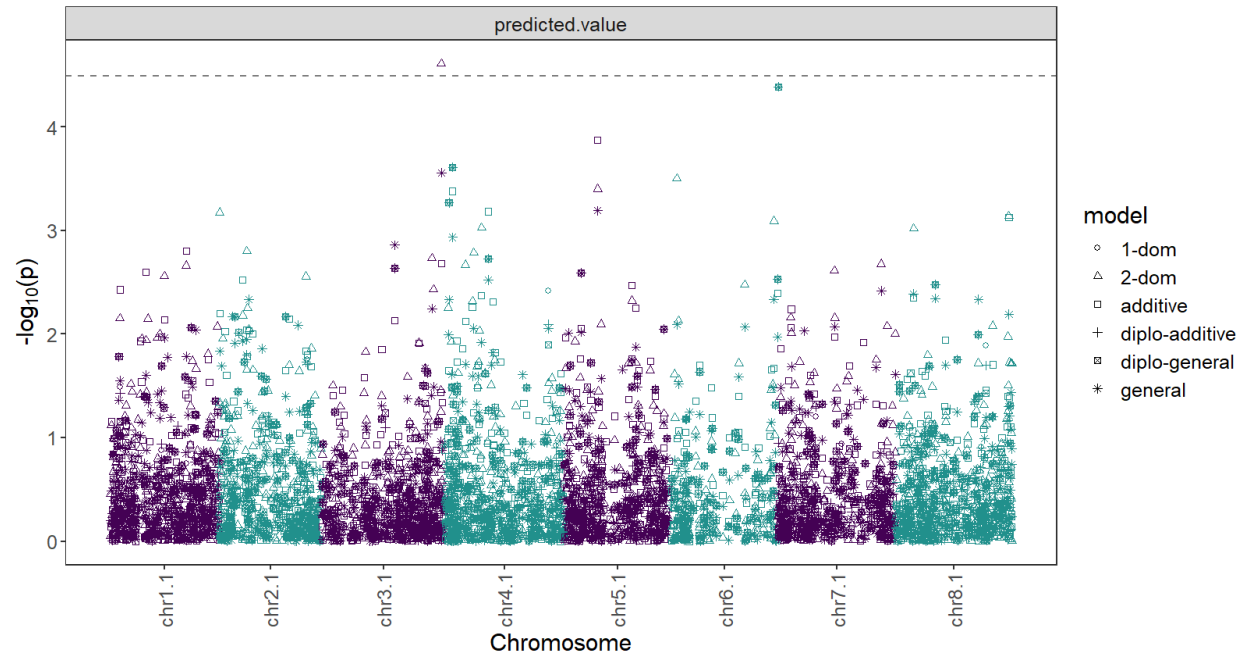


Figure 3. The Manhattan plot of GWAS model for plant height using overall harvest (ST2) BLUPs across 2,769 DarTag marker.

3.8 Markers Associated with Persistence and Regrowth Score

Fifteen unique markers were associated with persistence, and 16 unique markers were associated with regrowth score. Notably, all individual-harvest persistence associations were identified exclusively through dominance models (2-dom-alt and 2-dom-ref), with no associations detected under additive or diplo-additive models at the individual harvest level. This indicates a predominantly non-additive genetic architecture for persistence, consistent with the autotetraploid nature of alfalfa in which dominance and overdominance effects are common for complex survival traits. Shared markers between persistence and regrowth score included Chr1_1_040541523, Chr3_1_053349663, and Chr4_1_044553513 (glucomannan 4-beta-mannosyltransferase 2-like isoform X1).

Both traits shared markers with plant height (Chr3_1_090153084 annotated to omega-hydroxypalmitate O-feruloyl transferase-like; Chr2_1_003485139), and regrowth score shared Chr1_1_060661578 (Sorting nexin Vps5-like) with dry weight. Persistence and plant height shared Chr5_1_051075912. The most environmentally stable regrowth score marker was Chr3_1_090153084 (Chr3, omega-hydroxypalmitate O-feruloyl transferase-like), detected across five harvest environments (Davis3, Davis4, Prosser1, ST2_GENO, and ST2_GxE) with a consistently negative effect direction (−0.78 score units in Davis3, −0.68 in Davis4, −0.40 in Prosser1, −0.47 in ST2_GENO), suggesting a repeatable and environment-robust association suitable for marker-assisted selection. Additionally, Chr4_1_017342544 (Chr4) for regrowth score was detected across all five genetic models (general, additive, diplo-additive, diplo-general, and 1-dom-alt) in two independent Westside environments (Westside2 and Westside overall harvest), with a consistent positive effect (+1.28 score units in Westside2, +0.80 in Westside overall harvest), making it the most model-robust regrowth score marker in this study.

Additional persistence-specific coding markers included Chr1_073381901, annotated to CRM-domain containing factor *CFM3A* (chloroplastic/mitochondrial isoform X2); Chr2_007034650, annotated to Hexosyltransferase, *PMR5N* domain-containing protein, Mg-protoporphyrin IX chelatase, and *LRR* receptor-like kinase; and Chr3_005267797, annotated to F-box protein interaction domain protein and WPP domain-interacting protein 1. For regrowth score, Chr8_005343637 was annotated to Disease resistance protein (*TIR-NBS-LRR* class) and Chr3_087196712 to Superoxide dismutase (EC 1.15.1.1). A notable cross-trait marker was Chr4_044864682 (*ArnA*), Chr8_011413748 (*NBS-LRR*), Chr2_010055664 (*IRE1*), Chr3_070951996 (*PRA1*), and Chr7_024109724 (*GDPD6*), suggesting overlapping genetic architecture. Unique markers for persistence included annotations for phosphoenolpyruvate carboxylase (*PEPC*; Chr2_063400783) and a thaumatin-like protein (Chr3_068500656). For regrowth score, Chr3_062442788 was annotated to phosphatidate cytidyltransferase, Chr5_008387670 to malate dehydrogenase, and Chr8_028078352 to a *TGA2.2*-like transcription factor involved in salicylic acid signaling (Figure 4).

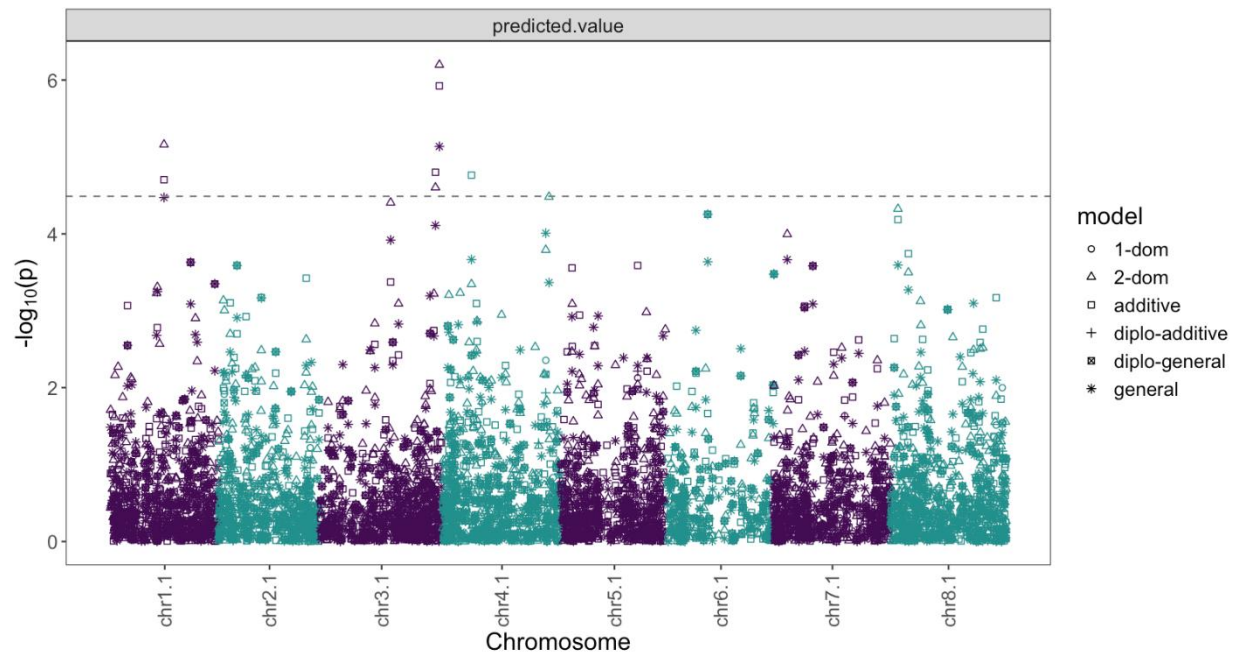


Figure 4. The Manhattan plot of GWAS model for regrowth score using overall harvest (ST2) BLUPs across 2,769 DarTag marker, the marker Chr7_024109724 was associated with *GDPD6* gene for regrowth

3.9 Genomic Prediction Accuracy

Genomic prediction was assessed using GBLUP with a five-fold cross-validation scheme (80% training / 20% testing, 100 iterations). For dry weight, predictive ability based on Pearson's correlation ranged from 0.17 (Five Points, WS2; Westside2) to 0.75 (Davis, Davis2) across harvest-specific single harvest models (Figure 5). The overall analysis of yield combined BLUP model yielded a predictive ability of 0.26. The lowest predictive ability at Westside2 is consistent with the lowest heritability recorded at that environment ($H^2 = 0.14$). With the exception of Westside2, Overall analysis of dry weight predictive ability was lower than those from most individual environments, likely reflecting the dilution of harvest-specific genetic signal when all environments are combined into a single BLUP.

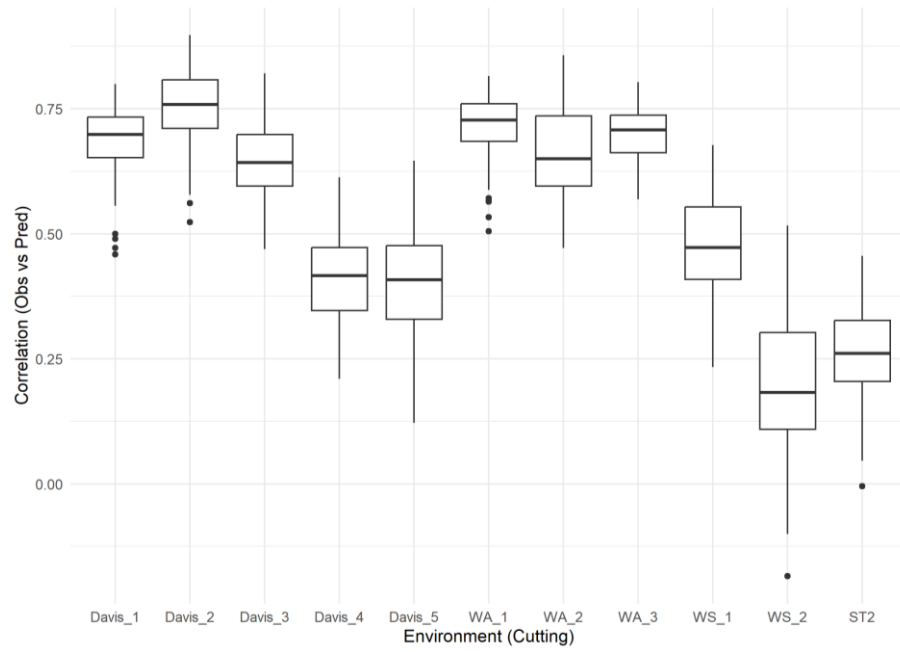


Figure 5. The boxplots of predicted ability of dry weight average by model by single harvest, BLUE (ST0) and overall stagewise BLUP (ST2) using 2,769 SNPs from 3k DArTag markers.

For plant height, predicted ability ranged from 0.54 to 0.84 across single harvest and overall plant height models (Figure 6). The highest predicted ability was obtained at the first Davis harvest in fall 2021 (Davis1, 0.84), and the lowest was at the overall plant height model (0.54). Notably, the pattern at Westside2 contrasts sharply with that observed for dry weight at the same location: while dry weight predicted ability was the lowest of all environments at Westside2 (WS2) (0.17), plant height predicted ability at Westside2 remained moderate and broadly consistent with other environments. This divergence indicates that the saline stress at Five Points (Westside) suppressed the genetic expression of biomass production but did not equivalently reduce the heritability or predictability of plant height, which is more constitutively determined. As with dry weight, the overall plant height predicted ability for plant height was lower than most harvest-specific predicted ability, indicating that environment-specific models retain more predictive signals than a single across-environment predictor.

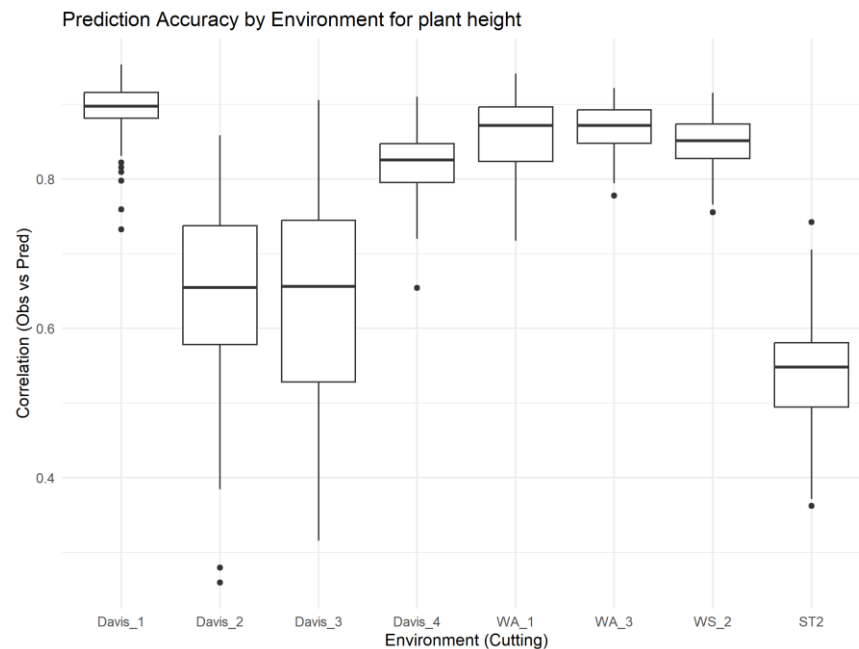


Figure 6. The boxplots of predicted ability (PA) of plant height average by model by single harvest, BLUE (ST0) and stagewise BLUP (ST2) using 2,769 SNPs from 3k DArTag markers

4. Discussion

4.1 Spatial variation and heritability

In large-scale multi location germplasm evaluations, spatial variation within and across field trials represents a major source of non-genetic phenotypic variation that must be accounted for before genotypic evaluation. In this study, the multi-environment trials data was first analyzed using a single harvest model, which returns the adjusted mean of accessions in each environment. The single harvest model analysis assumed accessions as fixed effects, blocks as random effects, and row and column positions as residuals, thereby incorporating spatial vectors directly into the experimental design. These spatial models correct the residual due to field variation, especially in imbalanced experimental designs or experiments with considerable missing data. The P-splines spatial model assumes uncorrelated variation between each entry value with independent covariance, making it well suited to capturing the gradual spatial trends that are typical of large field trials. The multi-environment trials study of alfalfa (Lin et al., 2023, Medina et al., 2025) confirmed that the P-splines spatial model is the most suitable model for multi-environment trials data analysis and the best fit for multi-environment alfalfa studies, supporting its use here as the primary method for obtaining spatially adjusted BLUEs. The low heritability at Five Points particularly Westside2 ($H^2 = 0.14$) is consistent with the saline irrigation water ($EC \sim 8.5 \text{ dS m}^{-1}$) used at that location, which likely imposed strong environmental stress that masked genotypic differences.

Obtaining spatially corrected means at the single-trial level is crucial before obtaining genotypic value across environments using multi-environment trial data. Multi-environment trials are necessary to determine the genetic merit of a genotype during genotype evaluation, which

requires modeling to obtain corrected means. Mixed linear models are one of the primary approaches to analyze multi-environment trials data: the benefits of using mixed linear models are the combination of random and fixed effects into a single model, which overcomes the issue of imbalanced or incomplete block designs and adjusts means for missing data by incorporating the heterogeneity of variance and covariance in the model (Isik et al., 2017). In this study, the multi-environment trials data from single-trial and multiple trials were analyzed using two mixed linear model approaches to determine the genotypic value of non-dormant alfalfa germplasm. The adjusted means, BLUEs and BLUPs are crucial values for determining genotype-by-environment interaction in terms of spatial and environmental variations (Smith et al., 2001).

The data revealed large differences in stand survival among accessions. These wide ranges suggest considerable genetic diversity in this collection that could be exploited for improvement of non-dormant alfalfa. The negative relationship between persistence and yield-related traits likely reflects a tradeoff in resource allocation: accessions that produce high biomass or grow tall may place greater demands on root carbohydrate reserves needed for stand survival through multiple harvest cycles (Cunningham et al., 1998). This was particularly true at Prosser where the cold winter likely selected against non-dormant material.

To combine information across the ten harvest time points and three locations, a stagewise analysis was performed using the spatially adjusted BLUEs as its foundation. BLUEs from single harvest and the variance-covariance matrix from each single harvest model were used as inputs for the overall harvest model. To determine the importance of genotype-by-environment effects (GxE), the Akaike Information Criterion (AIC) was used to compare a genotype main-effect model (GENO) with a GxE model for each trait. Only the GxE model for plant height showed a lower AIC, making it the best-fit model for that trait, while the GENO model showed better fit for dry weight,

persistence, and regrowth score. The difference in AIC between models indicates that the adjustment of mixed linear models is required to find the best-fit model based on the nature of each agronomic trait, and that the spatially corrected single harvest BLUEs provide a reliable phenotypic basis from which to estimate overall harvest BLUPs for use in GWAS and genomic prediction.

4.2 Linkage disequilibrium and marker-QTL detection

The ability of GWAS to detect marker-trait associations depends on the extent of linkage disequilibrium (LD) between genotyped markers and the true causal polymorphisms. LD describes the non-random association of alleles at two or more loci and determines how far a significant marker signal extends across the genome. When LD decays rapidly with physical distance, a marker must lie very close to a causal variant to capture its effect, requiring a much higher marker density for reliable detection (Flint-Garcia et al., 2003). Alfalfa is an autotetraploid with a large effective population size and a long history of recombination, all of which accelerate LD decay. Li et al. (2014) evaluated genome-wide LD in diverse tetraploid alfalfa using an Infinium SNP array and found that LD decayed to $r^2 = 0.20$ within approximately 26 kilobases across all genotypes, though this varied among subpopulations, LD extended to ~170 kilobases within narrow breeding populations and decayed within ~54 kilobases when parents from multiple origins were combined. More recently, using the 3K DArTag panel in a diverse alfalfa breeding population, LD was estimated to extend to only 8.575 megabases at $r^2 < 0.05$, reflecting the marker spacing that the DArTag platform (Sandercock et al., 2025). These contrasting estimates illustrate that LD in alfalfa is highly dependent on the population structure and genetic diversity of the panel used.

In this study, 383 accessions from diverse global origins were genotyped using 2,769 DArTag markers distributed across eight chromosomes, giving an average marker spacing of

several megabases. Given the rapid LD decay reported for diverse alfalfa germplasm (Li et al., 2014), this marker density is likely insufficient to ensure that every causal variant is in strong LD with at least one genotyped marker. Under the GWASpoly framework, GWAS relies on LD between markers and QTL for detection: even markers in the physical vicinity of a QTL may fail to reach statistical significance if allele frequency differences at the two loci reduce the LD between them (Rosyara et al., 2016). This is compounded by the fact that the study germplasm spans dormant, semi-dormant, and non-dormant accessions from multiple continents, a diversity that further accelerates LD decay and lowers the average r^2 between any given marker and nearby functional variants. The relatively low number of significant markers detected across harvests and environments is therefore consistent with a scenario in which the 2,769-marker panel captures only a fraction of the functional variation segregating in this diverse panel, rather than reflecting a simple genetic architecture for the traits. Increasing marker density, for example through higher-density sequencing-based approaches such as GBS, would be expected to substantially improve QTL detection power in germplasm collections (Hamblin et al., 2011).

The consequence of limited LD coverage affects the confidence with which a significant marker can be linked to a nearby gene. When LD decays over short distances, a significant marker is likely tagging a relatively small genomic window. Gene annotations reported for significant markers in this study were therefore assigned based on the physical position of the marker in the *Medicago sativa* reference genome using a flanking window approach, following the convention of previous alfalfa GWAS studies (Li et al., 2014; Medina et al., 2025). This approach identifies candidate genes but cannot confirm that the annotated gene is the causal locus; fine mapping would be required to establish causality. Nonetheless, the identification of biologically plausible gene annotations including cell wall synthesis enzymes, stress-response proteins, and transcription

factors at multiple associated markers suggests that the significant associations detected here are enriched for true positives rather than false discoveries driven by population structure.

4.3 Markers-traits association

The distribution of 60 unique significant markers (69 unique marker-trait associations) across all eight chromosomes suggests that dry weight, plant height, persistence, and regrowth score are polygenic traits. Markers were most numerous on Chromosome 1 (11 unique markers) and Chromosome 8 (10), followed by Chromosomes 2, 3, and 4 (8 each), highlighting these regions as harboring important QTL for agronomic traits in non-dormant alfalfa. This distribution warrants further investigation to benefit genomic selection for dry weight and plant height traits. The 24 marker-trait pairs detected in the overall harvest analysis can be divided into two functional classes: 20 that were unique to the overall harvest analysis (not detected in any individual harvest) likely represent small-effect, environmentally stable QTL whose signals are individually below significance but emerge through pooled statistical power; while the 4 pairs detected in both individual harvests and overall harvest represent the most robust associations, combining single-environment significance with multi-environment replication. The proportion of non-coding markers also varied substantially by trait, with persistence showing the highest ratio (10 non-coding vs 5 coding), suggesting that regulatory variation may be an important component of the genetic architecture of persistence. This has implications for marker-assisted selection, as non-coding markers may tag regulatory elements rather than functional protein variants, making the causal mechanism more difficult to resolve without additional functional genomics data.

The highest numbers of unique markers were on Chromosomes 1 (11) and 8 (10), suggesting these regions harbor important QTL for agronomic traits in non-dormant alfalfa,

warranting further fine-mapping efforts. Among the 24 marker-trait pairs detected in the overall harvest model analysis, 20 were unique to overall harvest (not detected in any individual harvest), suggesting they represent small-effect, environmentally stable QTL whose signals are individually below the significance threshold but emerge with pooled statistical power. The 5 marker-trait pairs detected in both individual harvests and overall harvest represent the most robust associations, combining single-environment significance with multi-environment replication.

This set of 60 unique markers does not fully overlap with markers identified for biomass and plant height in the previous studies (Yu, 2017; Lin et al., 2021; Lin et al., 2023). The explanation might be that all agronomic traits in this study were controlled by multiple loci with small effects, these loci could be affected by environment such as harvest or cutting time throughout the year and harvest x locations interaction with allele by harvest interaction (Robins et al., 2007). Or the interaction between phenotypic and locations which could affect the GWAS results that can be related to heritability at each harvest. The most recurrent marker was Chr6_1_080096025 (Chr6, position 80,096,025), annotated to a *TIR-NBS-LRR RCT1* resistance protein, detected across four harvest environments for dry weight (Davis1, Davis2, Davis3, and over all harvest ST2_GENO) and also significant for plant height, making it the strongest candidate for cross-environment and cross-trait follow-up. A second dry weight marker of high biological interest is Chr8_1_014266142 (Chr8), which produced the largest single-harvest additive effect in this study (+23.06 g plant⁻¹ in Davis1) and was detected under all five genetic models, indicating that its association is robust to model assumptions. Similarly, Chr4_1_017342544 (Chr4) was the most model-robust regrowth score marker, detected across five genetic models in two independent Westside environments with a consistent positive effect (+0.80–1.28 score units). The consistency

of effect direction across environments and models for these markers distinguishes them as priority candidates for follow-up in marker-assisted selection programs.

4.4 The overlapped marker and correlated traits

Chr4_044864682 which is associated with *ArnA* is a significant marker which is associated with all four traits. This protein plays an important role in synthesizing Lipopolysaccharide (LPS) which is one of the mechanisms to resist polymyxin in gram-negative pathogens (Gatzeva et al., 2005, Liu et al., 2025), while plants do not possess LPS, the annotation of Chr4_044864682 as an *ArnA*-like protein suggests a role in essential decarboxylases processes. *ArnA*-like homologs in plants are often involved in the interconversion of nucleotide sugars which is the primary substrates for cell wall composition (Bar-Peled & O'Neill, 2011), directly impact the structural phenotypes such as dry weight and plant height.

4.5 Key associated proteins and their function on agronomic traits.

The identification of E3 ubiquitin protein ligase (Chr1_066462122) and RING-type E3 ubiquitin transferase (Chr7_016949461) suggests that efficient protein degradation and recycling are vital for biomass accumulation. This is consistent with studies showing that protein quality control is essential during rapid growth stages (Eisele & Wolf, 2008). The study in *Arabidopsis thaliana* found that 6% of transcriptomes coding to ubiquitin protein ligase indicated their importance for protein quality control during growth and development (Vierstra, 2009)

The association of Phosphoenolpyruvate carboxylase (*PEPC*) (Chr2_063400783) with persistence and Malate dehydrogenase (Chr5_008387670) with regrowth score points toward metabolic efficiency. These two proteins are central to carbon metabolism; their role here is likely to support the rapid energy demands required for a plant to rapidly regrow after harvest (O'Leary

et al., 2011; Tomaz, T., et al. 2010). Effect sizes for persistence-associated markers ranged from approximately -1.15 to $+1.26$ plants per plot, while regrowth score effects ranged from approximately -1.03 to $+1.28$ score units, and plant height effects ranged from approximately -6.39 to $+10.58$ cm per plant. These moderate effect sizes are consistent with the polygenic architecture of these traits and suggest that multiple markers will need to be combined in a genomic selection index to achieve meaningful genetic gain.

The association of persistence and regrowth score *TGA2.2*-like transcription factors (Chr8_028078352) and Dual specificity protein phosphatase (*DSP8*) (Chr8_068994141) suggests that the plant's ability to modulate its internal whole plant signaling (Niggeweg et al., 2000), Absciscic acid signaling and stress response (Umezawa et al., 2009) is a prerequisite for high regrowth scores.

Chr2_010055664 which is associated with *IRE1* involved with managing cellular stress which is necessary for plant stability (Li & Howell, 2021) Chr1_016763798 associated with tRNA:m(4)X modification enzyme; *TRM13* which is responsible for specific methylation of tRNA, a modification to stabilize tRNA structure. This protein is essential for rapid and accurate protein synthesis. which is required during growth and development (Tkaczuk, 2010).

4.6 Markers with dry weight and plant height

Dry weight and plant height are very important traits for alfalfa because these two traits can determine the alfalfa variety with high yield and high biomass. Currently due to many reasons such as, long selection cycle, low breeding investment, out crossing with high breeding depression, and the lack of new germplasm diversity (Annicchiarico et al, 2015). In this study we found high variation of yield

across locations and harvests, indicating that the dry weight and plant height are controlled by small-effect loci, the same result was found in (Lin et al., 2023 and Medina et al., 2025) study.

In this study, marker Chr8_011413748 was located in the coding region of *TIR-NBS-LRR RCT1* resistance protein (Fragment) (Tan and Wu, 2022), which is an important gene for disease resistance in various crops including legume crops. This gene also plays an important role for abiotic stress (Chini et al., 2004). In alfalfa, *NBS-LRR* has been reported for biomass yield under drought stress (Yu 2017) and Neutral Detergent Fiber at 30 and 48 h (Lin et al., 2020), indicating that *NBS-LRR* proteins may be involved in multiple biological responses in alfalfa. The marker Chr8_1_014266142 (Chr8, position 14,266,142) was detected across all five genetic models for dry weight in Davis1 (Score = 6.31), with the largest single-harvest additive effect in this study (+23.06 g plant⁻¹) and represents a high-priority candidate for dry weight improvement in marker-assisted selection. Dry weight associations were predominantly positive in effect direction (15 positive vs. 1 negative among additive and 1-dom-alt effects), indicating that favorable alleles at most loci increase biomass. However, several loci showed large negative effects that breeders should select against, most notably Chr5_1_020147516 (Chr5; -38.1 g plant⁻¹ in the Prosser1 environment under diplo-additive and 1-dom-alt models) and Chr3_1_075963488 (Chr3; -99.8 plant⁻¹ in the Prosser2 environment under the 2-dom-alt model), the latter annotated to protein POLLEN DEFECTIVE IN GUIDANCE 1 isoform X2. The exceptionally large negative effect of Chr3_1_075963488 and its pollen guidance annotation may indicate a pleiotropic relationship between reproductive development and biomass allocation in alfalfa, a connection that warrants further functional investigation.

Chr3_090153084 was found on coding region of omega-hydroxypalmitate O-feruloyl transferase-like (Molina et al., 2009), this gene involved in suberin synthesis which is the important

substrate for plant cell wall synthesis and plant cell repair. This gene was annotated to *Medicago truncatula* and might play an important role in alfalfa growth.

4.7 Markers with persistence and regrowth score

Persistence is the number of plants per unit area remaining after a period of time, and regrowth score is a measurement of regrowth ability of alfalfa after cutting, Chr4_044553513 was found associated with these traits. This marker was annotated on glucomannan 4-beta-mannosyltransferase 2-like isoform X1 gene or *AtCSLA7* gene which involved pollen tube growth and embryogenesis in *Arabidopsis* (Goubet et al., 2003). This gene is involved in plant growth and toxin production in *Abrus precatorius* (Hovde et al., 2019). This may lead to finding genes that have an association between persistence and regrowth score, however there is no common marker between dry weight and regrowth score from this study.

Chr8_072762376 annotated to Transportin-3 and protein JINGUBANG-like isoform X1, these genes play an important role for RNA splicing and plant development. (Shin et al., 2019) Chr3_005267797 annotated with F-box protein interaction domain protein and WPP domain-interacting protein 1 (Young et al., 2011) involved in root development in *Medicago truncatula*.

For regrowth score, Chr8_005343637 has been annotated to disease resistance protein (*TIR-NBS-LRR* class) (Tan and Wu, 2022), who also found association with marker Chr8_011413748 for dry weight. Chr3_090153084 (omega-hydroxypalmitate O-feruloyl transferase-like) and Chr4_044553513 (glucomannan 4-beta-mannosyltransferase 2-like isoform X1) have been annotated to the plant cell wall synthesis genes (Elbein, 1969; Gou et al., 2009) and Chr3_087196712 was annotated to Superoxide dismutase (EC 1.15.1.1), this gene protects cells from free radicals oxidative stress. (Cai et al., 2019) persistence and regrowth are the important traits

that reflect the persistence of the alfalfa stand after multiple harvests, in this study, we found multiple genes that may be involved in plant germination and plant development for these two traits.

4.8 Genomic Prediction

Predictive ability varied considerably across environments and traits in this study, a pattern that is well established in the genomic prediction literature and is primarily explained by differences in heritability among environments. For dry weight, predictive ability ranged from 0.17 at Westside (WS2) to 0.75 at Davis (Davis2) across harvest-specific single harvest models, while the across-environment overall harvest model yielded a predictive ability of 0.26. For plant height, predictive ability ranged from 0.54 to 0.84 across single harvest and overall models. The theoretical relationship between heritability and predictive ability is well documented, genomic prediction accuracy is bounded by the square root of heritability, so environments with low heritability necessarily yield lower predictive ability (Daetwyler et al., 2010; Wientjes et al., 2013). This was most clearly illustrated at Westside2, where saline irrigation water (EC ~8.5 dS m⁻¹) suppressed dry weight heritability to $H^2 = 0.14$, the lowest recorded across all environments, and produced the lowest predictive ability of 0.17. In contrast, plant height at the same location retained moderate predictive ability broadly consistent with other environments, because this trait is more constitutively determined and its heritability was less sensitive to the abiotic stress imposed at that site. These results are consistent with findings in other forage and field crops showing that the reliability of training phenotypes is the primary driver of environment-to-environment variation in genomic prediction accuracy (Combs & Bernardo, 2013).

Genomic prediction is especially valuable for traits like dry weight that are controlled by a large number of loci each with small individual effects, because such polygenic architectures are difficult to

exploit through marker-assisted selection based on individual QTL. Previous studies in alfalfa have demonstrated how marker panel composition affects predictive ability: Medina et al. (2025) performed GBLUP using DArTag markers selected from a GWAS model and achieved a predictive ability of up to 90% for overall yield, while Zhang et al. (2023) used the top 3,000 GWAS markers to obtain a predictive ability of 64% for fall dormancy. Both approaches relied on pre-selected, trait-associated markers to maximize signal, whereas the present study used the full 2,769-marker DArTag panel without GWAS-based pre-selection, making the predictive ability values obtained here more conservative but also more reflective of a routine breeding scenario.

A key consideration in marker-based genomic prediction is the risk of overfitting when markers are selected based on association with training phenotypes. The 3K DArTag marker platform was specifically designed to mitigate this risk by capturing a representative set of markers distributed across the entire alfalfa genome, selected from a diverse panel of germplasm, rather than being optimized for any single population or trait (Zhao et al., 2023). Its broad genomic coverage and high reproducibility make it a candidate genotyping platform for routine alfalfa breeding, and ten-fold cross-validation with 100 iterations (90% training / 10% testing) was used to obtain stable PA estimates under realistic conditions (Sipowicz et al., 2025).

In this study, the overall harvest model using 2,769 DArTag markers yielded a PA of 0.26 for dry weight, which was lower than most single harvest predictive ability values. This pattern likely reflects the dilution of harvest-specific genetic signal when environments are combined into a single across-environment BLUP and is consistent with the general observation that environment-specific models retain more predictive signal than a single across-environment predictor (Burgueño et al., 2012). Nonetheless, the predictive ability obtained with an unselected genome-wide marker panel and spatially corrected phenotypes demonstrates that the combination

of the 3K DArTag platform with rigorous mixed-model phenotypic analysis is a cost-effective foundation for genomic selection in alfalfa breeding programs (Sipowicz et al., 2025). Increasing marker density, expanding training population size, or incorporating environment-specific prediction models could further improve predictive ability, particularly in high-stress environments such as Westside where heritability constrains the ceiling of achievable accuracy.

5. Conclusion

The study integrated the two approaches to perform GWAS and genomic prediction, first approach, using stagewise analysis to estimate the genotype by environment effect and detected to spatial variation, obtain adjusted phenotypic mean of accession and using highly reproducible and cost effective, the 3k DArTag platform that developed from diverse alfalfa germplasm to increase genotyping quality. In this study, we analyzed four agronomic traits, dry weight, plant height, persistence and regrowth score using two statistical models to obtain the adjusted means in the form of BLUEs and BLUPs. The BLUEs and BLUPs values were used in GWAS and identified sixty-nine markers across all harvests. This workflow can find the marker that is associated in multiple harvests and traits. The DArTag genotype matrix was used as an input along with BLUPs for genomic prediction for dry weight. The predictive ability was different based on the harvest, stagewise and DArTag markers approach can increase prediction accuracy of genomic prediction models. This study shows the importance of using both phenotypic and genotypic approaches to improve the alfalfa genomic study. These two approaches could be used for alfalfa molecular breeding to improve alfalfa agronomic traits in breeding programs.

6. Acknowledgement

This research was funded by grants from the USDA-Alfalfa Seed and Alfalfa Forage Research Program #Award Number 2019-70005-30361 to ECB and BI, from the California Crop Improvement Association to ECB, by Breeding Insight, and by USDA-ARS. Kreingkrai Nonkum is funded by a Royal Thai Government Scholarship. This project is part of Hatch Multistate Research Project NE-2210 (formerly NE1710).

7. References

- Amadeu RR, Garcia AA, Munoz PR, and Ferrão LF. (2023). AGHmatrix: genetic relationship matrices in R. *Bioinformatics*. Jul 1;39(7):btad445. <https://doi.org/10.1093/bioinformatics/btad445>
- Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second*. NY: Springer-Verlag, 63(2020), 10.
- Annicchiarico P, Barrett B, Brummer EC et al (2015). Achievements and challenges in improving temperate perennial forage legumes. *CRC Crit Rev Plant Sci* 34:327–380. <https://doi.org/10.1080/07352689.2014.898462>
- Bateman A (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47:D506–D515. <https://doi.org/10.1093/nar/gky1049>
- Bar-Peled, M., & O'Neill, M. A. (2011). Plant nucleotide sugar formation, interconversion, and salvage by sugar recycling. *Annual review of plant biology*, 62(1), 127-155.
- Blondon, F., Marie, D., Brown, S. and Kondorosi, A. (1994). Genome size and base composition in *Medicago sativa* and *M. truncatula* species. *Genome* doi:10.1139/g94-037.

- Brummer, E. C., & Casler, M. D. (2004). Genomics research in alfalfa. *In* R. F. Wilson, H. T. Stalker, & E. C. Brummer (Eds.), *Legume Crop Genomics* (pp. 110–142). AOCS Press.
- Genomics research in alfalfa
- Brummer E.C. (2002). Agronomic consequences of dormant-nondormant alfalfa mixtures. *Agron. J.* 94 782-785 <http://doi.org/10.2134/agronj2002.0782>
- Brummer, E. Charles, M. Maroof Shah, and Diane Luth. (2000). "Reexamining the relationship between fall dormancy and winter hardiness in alfalfa." *Crop Science* 40, no. 4 : 971-977.
- Brummer, E. C. (1999). "Capturing heterosis in forage crop cultivar development." *Crop science* 39, no. 4 : 943-954.
- Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J. (2012). Genomic Prediction of Breeding Values when Modeling Genotype $\times$ Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Science*, 52: 707-719.
- <https://doi.org/10.2135/cropsci2011.06.0299>
- Byrne, P. F., Volk, G. M., Gardner, C., Gore, M. A., Simon, P. W., Smith, S. (2018). Sustaining the future of plant breeding: The critical role of the USDA-ARS National Plant Germplasm System. *Crop Sci.* 58, 451–468. doi: 10.2135/cropsci2017.05.0303
- <https://access.onlinelibrary.wiley.com/doi/full/10.2135/cropsci2017.05.0303>
- Castonguay, Y., Laberge, S., Brummer, E. C., & Volenec, J. J. (2006). Alfalfa winter hardiness: a research retrospective and integrated perspective. *Adv. Agron.*, 90, 203-265.
- Cai Y, Cai X, Wang Q, Wang P, Zhang Y, Cai C, Xu Y, Wang K, Zhou Z, Wang C, Geng S, Li B, Dong Q, Hou Y, Wang H, Ai P, Liu Z, Yi F, Sun M, An G, Cheng J, Zhang Y, Shi Q,

- Xie Y, Shi X, Chang Y, Huang F, Chen Y, Hong S, Mi L, Sun Q, Zhang L, Zhou B, Peng R, Zhang X, Liu F. (2020). Genome sequencing of the Australian wild diploid species *Gossypium australe* highlights disease resistance and delayed gland morphogenesis. *Plant Biotechnol J*. 2020 Mar;18(3):814-828. doi: 10.1111/pbi.13249. Epub 2019 Sep 13. PMID: 31479566; PMCID: PMC7004908.
- Chen, H.T. , Zeng, Y. , Yang, Y.Z. , Huang, L.L. , Tang, B.L. , Zhang, H. , Hao, F. *et al.* (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* 11, 2494.
- Chini, A., Grant, J.J., Seki, M., Shinozaki, K. and Loake, G.J. (2004). Drought tolerance established by enhanced expression of the *CC-NBS-LRR* gene, *ADRI*, requires salicylic acid, EDS1 and ABI1. *The Plant Journal*, 38: 810-822. <https://doi.org/10.1111/j.1365-3113.2004.02086.x>
- Clark LV, Lipka AE, and Sacks EJ (2019). polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3* 9(3) 663--673. doi:[10.1534/g3.118.200913](https://doi.org/10.1534/g3.118.200913)
- Combs, E. and Bernardo, R. (2013). Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. *The Plant Genome*, 6: plantgenome2012.11.0030. <https://doi.org/10.3835/plantgenome2012.11.0030>
- Covarrubias-Pazaran, G., (2025). lme4breeding: a fast linear mixed model for multi-trait/multi-environment experiments. *bioRxiv* 2024.05.17.594795; doi: <https://doi.org/10.1101/2024.05.17.594795>

Covarrubias-Pazaran, G., (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PloS one*, 11(6), p.e0156744.

Cunningham, S.M., Volenec, J.J. and Teuber, L.R. (1998), Plant Survival and Root and Bud Composition of Alfalfa Populations Selected for Contrasting Fall Dormancy. *Crop Science*, 38: 962-969

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*. 2010 Jul;185(3):1021-31. doi: 10.1534/genetics.110.116855. Epub 2010 Apr 20. PMID: 20407128; PMCID: PMC2907189.

Damesa TM, Möhring J, Worku M, Piepho H (2017). One step at a time: stage-wise analysis of a series of experiments. *Agron J* 109:845–857. <https://doi.org/10.2134/agronj2016.07.0395>

Eisele, F., & Wolf, D. H. (2008). Degradation of misfolded protein in the cytoplasm is mediated by the ubiquitin ligase Ubr1. *FEBS letters*, 582(30), 4143-4146.

Elbein AD. Biosynthesis of a cell wall glucomannan in mung bean seedlings. (1969). *J Biol Chem*. 1969 Mar 25;244(6):1608-16. PMID: 4304230.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, 6(5), e19379.

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual review of plant biology*, 54(1), 357-374.

Francis, R. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Mol eco res*, 17(1), 27-32.

Gitanshu Munjal, Jingjie Hao, Larry R Teuber, E Charles Brummer, (2018). Selection Mapping Identifies Loci Underpinning Autumn Dormancy in Alfalfa (*Medicago sativa*), *G3*, Volume 8, Issue 2, 1 February 2018, Pages 461–468, <https://doi.org/10.1534/g3.117.300099>

Gatzeva-Topalova PZ, May AP, Sousa MC. (2005). Structure and mechanism of ArnA: conformational change implies ordered dehydrogenase mechanism in key enzyme for polymyxin resistance. *Structure*. 2005 Jun;13(6):929-42. doi: 10.1016/j.str.2005.03.018. PMID: 15939024; PMCID: PMC2997725.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440-1450.

Gou JY, Yu XH, Liu CJ. A (2009). hydroxycinnamoyltransferase responsible for synthesizing suberin aromatics in Arabidopsis. *Proc Natl Acad Sci U S A*. 2009 Nov 3;106(44):18855-60. doi: 10.1073/pnas.0905555106. Epub 2009 Oct 21. PMID: 19846769; PMCID: PMC2773987.

Goubet F, Misrahi A, Park SK, Zhang Z, Twell D, Dupree P. (2003). AtCSLA7, a cellulose synthase-like putative glycosyltransferase, is important for pollen tube growth and embryogenesis in Arabidopsis. *Plant Physiol*. 2003 Feb;131(2):547-57. doi: 10.1104/pp.014555. PMID: 12586879; PMCID: PMC166831.

- Govindaraj, M., Vetriventhan, M., & Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genetics research international*, 2015(1), 431487.: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383386/>
- Hamblin, M. T., Buckler, E. S., & Jannink, J. L. (2011). Population genetics of genomics-based crop improvement methods. *Trends in genetics*, 27(3), 98-106.
- Hovde BT, Daligault HE, Hanschen ER, Kunde YA, Johnson MB, Starkenburg SR, Johnson SL. (2019). Detection of Abrin-Like and Prepropulchellin-Like Toxin Genes and Transcripts Using Whole Genome Sequencing and Full-Length Transcript Sequencing of *Abrus precatorius*. *Toxins (Basel)*. 2019 Nov 25;11(12):691. doi: 10.3390/toxins11120691. PMID: 31775284; PMCID: PMC6950105.
- Isik F, Holland J, Maltecca C (2017). Multi environmental trials. *In*: Isik F, Holland J, Maltecca C (eds) Genetic data analysis for plant and animal breeding. Springer, Cham, pp 227–262
- Lesk, C., Rowhani, P., & Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. *Nature*, 529(7584), 84-87.
- Li X, Han Y, Wei Y, Acharya A, Farmer AD, Ho J, et al. (2014). Development of an Alfalfa SNP Array and Its Use to Evaluate Patterns of Population Structure and Linkage Disequilibrium. *PLoS ONE* 9(1): e84329. <https://doi.org/10.1371/journal.pone.0084329>
- Li, Z., & Howell, S. H. (2021). The two faces of IRE1 and their role in protecting plants from stress. *Plant Science*, 303, 110758.

- Lin, S., Medina, C.A., Boge, B. *et al.* (2020). Identification of genetic loci associated with forage quality in response to water deficit in autotetraploid alfalfa (*Medicago sativa* L.). *BMC Plant Biol* 20, 303 (2020). <https://doi.org/10.1186/s12870-020-02520-2>
- Lin S, Medina CA, Norberg OS *et al* (2021). Genome-wide association studies identifying multiple loci associated with alfalfa forage quality. *Front Plant Sci* 12:1–15. <https://doi.org/10.3389/fpls.2021.648192>
- Lin S, Medina CA, Wang G, Combs D, Shewmaker G, Fransen S, Llewellyn D, Norberg S, Yu LX. (2023). Identification of genetic loci associated with five agronomic traits in alfalfa using multi-environment trials. *Theor Appl Genet.* 2023 Apr 29;136(5):121. doi: 10.1007/s00122-023-04364-4. PMID: 37119337.
- Liu X, Yang R, Ren L, Li T, Li Y, Yan Z, Gao Y, Yang M, Li J. (2025). Structural Basis for Targeting the Bifunctional Enzyme *ArnA*. *Biomolecules.* 2025 Nov 13;15(11):1594. doi: 10.3390/biom15111594. PMID: 41301511; PMCID: PMC12650200.
- Long, R. , Zhang, F. , Zhang, Z. , Li, M. , Chen, L. , Wang, X. , Liu, W. *et al.* (2022). Genome assembly of alfalfa cultivar zhongmu-4 and identification of SNPs associated with agronomic traits. *Genomics Proteomics Bioinformatics*, 20, 14–28.
- Malosetti M, Ribaut J-M and van Eeuwijk FA (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* 4:44. doi: 10.3389/fphys.2013.00044

- Medina CA, Samac DA, Yu L-X (2021). Pan-transcriptome identifying master genes and regulation network in response to drought and salt stresses in Alfalfa (*Medicago sativa* L.). *Sci Rep* 11:17203. <https://doi.org/10.1038/s41598-021-96712-x>
- Medina, Cesar A., Dongyan Zhao, Meng Lin, Manoj Sapkota, Alexander M. Sandercock, Craig T. Beil, Moira J. Sheehan et al. (2024). "Pre-breeding in alfalfa germplasm develops highly differentiated populations, as revealed by genome-wide microhaplotype markers." *Sci Rep* 15, 1253 <https://doi.org/10.1038/s41598-024-84262-x>
- Medina CA, Heuschele DJ, Zhao D, Lin M, Beil CT, Sheehan MJ and Xu Z (2024). Multi-trait modeling and machine learning discover new markers associated with stem traits in alfalfa. *Front. Plant Sci.* 15:1429976. doi: 10.3389/fpls.2024.1429976\
- Medina, Cesar A., Julie Hansen, Jamie Crawford, Donald Viands, Manoj Sapkota, Zhanyou Xu, Michael D. Peel, and Long-Xi Yu. (2025). "Genome-Wide Association and Genomic Prediction of Alfalfa (*Medicago sativa* L.) Biomass Yield Under Drought Stress" *International Journal of Molecular Sciences* 26, no. 2: 608. <https://doi.org/10.3390/ijms26020608>
- Meirmans, P. G. (2020). GENODIVE version 3.0: easy-to-use software for the analysis of genetic data of diploids and polyploids. *Mol Eco Res.* 20, 1126–1131 (2020).
- Möhring J, Piepho H-P (2009). Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Sci* 49:1977–1988. <https://doi.org/10.2135/cropsci2009.02.0083>

- Molina I, Li-Beisson Y, Beisson F, Ohlrogge JB, Pollard M. (2009). Identification of an *Arabidopsis* feruloyl-coenzyme A transferase required for suberin synthesis. *Plant Physiol.* 2009 Nov;151(3):1317-28. doi: 10.1104/pp.109.144907. Epub 2009 Sep 16. PMID: 19759341; PMCID: PMC2773081.
- Mrode, R. A. (2005). *Linear models for the prediction of animal breeding values* (No. Ed. 2, pp. xiii+-344).
- NASS, USDA (2025). USDA quickstats. In: USDA National Agricultural Statistics Service. <https://quickstats.nass.usda.gov/> Accessed 20 July 2025
- Niggeweg, R., Thurow, C., Kegler, C., & Gatz, C. (2000). Tobacco transcription factor TGA2. 2 is the main component of as-1-binding factor ASF-1 and is involved in salicylic acid-and auxin-inducible expression of as-1-containing target promoters. *Journal of Biological Chemistry*, 275(26), 19897-19905.
- O'Leary, Brendan, Joonho Park, William C. Plaxton; (2011). The remarkable diversity of plant PEPC (phosphoenolpyruvate carboxylase): recent insights into the physiological functions and post-translational controls of non-photosynthetic PEPCs. *Biochem J* 15 May 2011; 436 (1): 15–34. doi: <https://doi.org/10.1042/BJ20110078>
- Pégard M, Barre P, Delaunay S, Surault F, Karagić D, Milić D, Zorić M, Ruttink T and Julier B (2023). Genome-wide genotyping data renew knowledge on genetic diversity of a worldwide alfalfa collection and give insights on genetic control of phenology traits. *Front. Plant Sci.* 14:1196134. doi: 10.3389/fpls.2023.1196134
- Pritchard, J. K., Wen, W., and Falush, D. (2003, July). *Documentation for STRUCTURE software: Version 2.*

Postman, J., Hummer, K., Stover, E., Krueger, R., Forsline, P., Grauke, L. J., et al. (2006). Fruit and nut genebanks in the US National Plant Germplasm System. *HortScience* 41, 1188–1194. doi: 10.21273/HORTSCI.41.5.1188

<https://www.academia.edu/download/102826888/article-p1188.pdf>

Putnam, D., Meccage, E. (2022). Profitable alfalfa production sustains the environment in *Proceeding 2022 World alfalfa congress* (Davis, CA: University of California Davis), 14–17. : <https://alfalfasyposium.ucdavis.edu/+symposium/proceedings/2022/Articles/Daniel-Putnam-2022-Profitable-Alfalfa-Production-Sustains-The-Environment.pdf>

Robins JG, Bauchan GR, Brummer EC (2007). Genetic mapping forage yield, plant height, and regrowth at multiple harvests in tetraploid alfalfa (*Medicago sativa* L.). *Crop Sci* 47:11–18. <https://doi.org/10.2135/cropsci2006.07.0447>

Rodríguez -Alvarez, M. X., Boer, M. P., van Eeuwijk, F. A., & Eilers, P. H. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, 23, 52-71.

Rosyara, U.R., De Jong, W.S., Douches, D.S. and Endelman, J.B., (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *The plant genome*, 9(2), pp.plantgenome2015-08.

- Sandercock, A. M. , Peel, M. D. , Taniguti, C. H. , Chinchilla-Vargas, J. , Chen, S. , Sapkota, M. , Lin, M. , Zhao, D. , Ackerman, A. J. , Basnet, B. R. , Beil, C. T. , & Sheehan, M. J. (2025). BIGapp: A user-friendly genomic tool kit identified quantitative trait loci for creeping rootedness in alfalfa (*Medicago sativa* L.). *The Plant Genome*, 18, e70067. 10.1002/tpg2.70067
- Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components*. John Wiley & Sons.
- Shen, C. , Du, H. , Chen, Z. , Lu, H. , Zhu, F. , Chen, H. , Meng, X. *et al.* (2020). The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Mol. Plant*, 13, 1250–1261.
- Shin, AY., Koo, N., Kim, S. *et al.* (2019). Draft genome sequences of two oriental melons, *Cucumis melo* L. var. *makuwa*. *Sci Data* 6, 220 <https://doi.org/10.1038/s41597-019-0244-x>
- Simmonds, N. W. Introgression and incorporation. Strategies for the use of crop genetic resources. *Biological Reviews of the Cambridge Philosophical Society* 68, 539–562 (1993). : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.1993.tb01243.x>
- Sipowicz, P., Murad Leite Andrade, M. H., Fernandes Filho, C. C., Benevenuto, J., Muñoz, P., Ferrão, L. F. V., Resende, M. F. R., Messina, C., & Rios, E. F. (2025). Optimization of high-throughput marker systems for genomic prediction in alfalfa family bulks. *The Plant Genome*, 18, e20526. <https://doi.org/10.1002/tpg2.20526>
- Smith A, Cullis B, Thompson R (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147. <https://doi.org/10.1111/j.0006-341X.2001.01138.x>

- Smith A.B., Ganesalingam A., Kuchel H., and Cullis B.R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theor Appl Genet* 128:55–72. <https://doi.org/10.1007/s00122-014-2412-x>
- Tan, S., and Wu, S. (2012). Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon*. *Int J. of Genom*, 2012(1), 418208.
- Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G, Wu S, Wang Y. (2023). SRplot: A free online platform for data visualization and graphing. *PLoS One*. 2023 Nov 9;18(11):e0294236. doi: 10.1371/journal.pone.0294236. PMID: 37943830.
- Tang F, Yang S, Gao M, Zhu H. (2013). Alternative splicing is required for RCT1-mediated disease resistance in *Medicago truncatula*. *Plant Mol Biol*. 2013 Jul;82(4-5):367-74. doi: 10.1007/s11103-013-0068-6. Epub 2013 May 9. PMID: 23657790.
- Teuber, L.R., K.L. Taggard, L.K. Gibbs, M.H. McCaslin, M.A. Peterson, and D.K. Barnes. (1998). Fall dormancy. In Standard tests to characterize alfalfa varieties. 3rd ed. (Amended 1998). North Am. Alfalfa Improvement Conf., Beltsville, MD
- Tkaczuk, K.L. (2010). Trm13p, the tRNA:Xm4 modification enzyme from *Saccharomyces cerevisiae* is a member of the Rossmann-fold MTase superfamily: prediction of structure and active site. *J Mol Model* 16, 599–606 (2010). <https://doi.org/10.1007/s00894-009-0570-6>

Tomaz Tiago, Matthieu Bagard, Itsara Pracharoenwattana, Pernilla Lindén, Chun Pong Lee, Adam J. Carroll, Elke Ströher, Steven M. Smith, Per Gardeström, A. Harvey Millar, (2010). Mitochondrial Malate Dehydrogenase Lowers Leaf Respiration and Alters Photorespiration and Plant Growth in Arabidopsis, *Plant Physiology*, Volume 154, Issue 3, November 2010, Pages 1143–1157, <https://doi.org/10.1104/pp.110.161612>

Umezawa, T., N. Sugiyama, M. Mizoguchi, S. Hayashi, F. Myouga, K. Yamaguchi-Shinozaki, Y. Ishihama, T. Hirayama, & K. Shinozaki, (2009). Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in Arabidopsis, *Proc. Natl. Acad. Sci. U.S.A.* 106 (41) 17588-17593, <https://doi.org/10.1073/pnas.0907095106>

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91, 4414–4423.

Vierstra, R. (2009). The ubiquitin–26S proteasome system at the nexus of plant biology. *Nat Rev Mol Cell Biol* **10**, 385–397 <https://doi.org/10.1038/nrm2688>

Weishaar, Mindy A., E. Charles Brummer, Jeffrey J. Volenec, Kenneth J. Moore, and Suzanne Cunningham. (2005). "Improving winter hardiness in nondormant alfalfa germplasm." *Crop science* 45, no. 1 : 60-65.

Wientjes YC, Veerkamp RF, Calus MP. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*. 2013 Feb;193(2):621-31. doi: 10.1534/genetics.112.146290. Epub 2012 Dec 24. PMID: 23267052; PMCID: PMC3567749.

Young ND, Debellé F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KA, Tang H, Rombauts S, Zhao PX, Zhou P, Barbe V, Bardou P, Bechner M, Bellec A, Berger A, Bergès H, Bidwell S, Bisseling T, Choisine N, Couloux A, Denny R, Deshpande S, Dai X, Doyle JJ, Dudez AM, Farmer AD, Fouteau S, Franken C, Gibelin C, Gish J, Goldstein S, González AJ, Green PJ, Hallab A, Hartog M, Hua A, Humphray SJ, Jeong DH, Jing Y, Jöcker A, Kenton SM, Kim DJ, Klee K, Lai H, Lang C, Lin S, Macmil SL, Magdelenat G, Matthews L, McCorrison J, Monaghan EL, Mun JH, Najar FZ, Nicholson C, Noirot C, O'Bleness M, Paule CR, Poulain J, Prion F, Qin B, Qu C, Retzel EF, Riddle C, Sallet E, Samain S, Samson N, Sanders I, Saurat O, Scarpelli C, Schiex T, Segurens B, Severin AJ, Sherrier DJ, Shi R, Sims S, Singer SR, Sinharoy S, Sterck L, Viollet A, Wang BB, Wang K, Wang M, Wang X, Warfsmann J, Weissenbach J, White DD, White JD, Wiley GB, Wincker P, Xing Y, Yang L, Yao Z, Ying F, Zhai J, Zhou L, Zuber A, Dénarié J, Dixon RA, May GD, Schwartz DC, Rogers J, Quétier F, Town CD, Roe BA. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*. 2011 Nov 16;480(7378):520-4. doi: 10.1038/nature10625. PMID: 22089132; PMCID: PMC3272368.

Yu LX. (2017). Identification of single-nucleotide polymorphic loci associated with biomass yield under water deficit in alfalfa (*Medicago sativa* L.) using genome-wide sequencing and association mapping. *Front Plant Sci* 8:1–11. <https://doi.org/10.3389/fpls.2017.01152>

Zhang, F., Kang, J., Long, R., Li, M., Sun, Y., He, F., Jiang, X., Yang, C., Yang, X., Kong, J. and Wang, Y., 2023. (2023). Application of machine learning to explore the genomic prediction accuracy of fall dormancy in autotetraploid alfalfa. *Hortic. Res.* 2023, *10*(1), p.uhac225.

Zhao, Dongyan, Katherine Maria Mejia-Guerra, Marcelo Mollinari, Deborah Samac, Brian Irish, Katarzyna Heller-Uszynska, Craig Thomas Beil, and Moira Jane Sheehan. (2023). "A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.)." In *Genetic resources*, vol. 4, no. 8, pp. 55-63.

Supplementary table 1. Significant marker-trait associations identified by GWAS across all genetic models and harvest environments.

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation	
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom		
Dry weight	Chr1_006491042	Chr1	6,491,042	Davis2	4.07	–	–	–	5.87	5.87	–	F-box/LRR-repeat protein 15 isoform X3; PABS domain-containing protein;thermospermine synthase ACAULIS5;hypothetical protein	
	Chr1_009325684	Chr1	9,325,684	Prosser1	4.45	–	–	–	–	–	–	–	
	Chr1_010654139	Chr1	10,654,139	Prosser	4.15	–	–	–	–	–	6.52	–	
	Chr1_030458048	Chr1	30,458,048	Prosser1	4.45	–	–	–	–	–	–	–	
	Chr1_060661578	Chr1	60,661,578	Davis4	4.07	–	–	–	21.43	21.43	–	Sorting nexin Vps5-like	
	Chr1_075444453	Chr1	75,444,453	Davis1	4.09	–	–	–	–	27.14	–	PHD finger protein At1g33420-like; testis-expressed protein 2 isoform X1; mavicyanin-like; protein yippee-like At4g27745	
	Chr2_043201762	Chr2	43,201,762	Westside1	4.45	–	–	–	–	–	–	–	

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom	
	Chr3_066254938	Chr3	66,254,938	Westside1	4.45	—	—	—	—	—	—	—
	Chr3_075963488	Chr3	75,963,488	Prosser2	4.27	—	—	—	—	—	-99.78	protein POLLEN DEFECTIVE IN GUIDANCE 1 isoform X2
	Chr3_080980552	Chr3	80,980,552	ST2_GENO	4.18	—	—	—	—	—	0.06	PALP domain-containing protein
	Chr3_088572986	Chr3	88,572,986	Westside2	4.15	—	—	—	—	—	12.53	—
	Chr4_004463235	Chr4	4,463,235	Prosser1	4.18	—	—	—	—	—	—	boron transporter 4 isoform X1
	Chr4_061541635	Chr4	61,541,635	ST2_GENO	4.32	—	—	—	—	—	0.03	—
	Chr4_064925789	Chr4	64,925,789	Prosser1	4.45	—	—	—	—	—	—	cyclin-B1-2-like
	Chr5_005193449	Chr5	5,193,449	ST2_GENO	4.32	—	—	—	—	—	-0.05	probable sodium/metabolite cotransporter BASS6, chloroplastic isoform X1
	Chr5_005988357	Chr5	5,988,357	Westside1	4.18	—	—	—	—	—	—	probable aminotransferase ACS10 isoform X3;probable

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation	
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom		
												aminotransferase ACS10 isoform X2.	
	Chr5_020147516	Chr5	20,147,516	Prosser1	4.09	–	–	–	-38.11	-38.11	–	–	
	Chr5_074362899	Chr5	74,362,899	ST2_GENO	3.45	–	–	–	–	-0.03	–	–	C2 domain-containing protein (Fragment)
	Chr6_080096025	Chr6	80,096,025	Davis	4.09	–	–	8.79	8.79	8.79	–	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr6_080096025	Chr6	80,096,025	Davis1	4.09	–	–	17.42	17.42	17.42	–	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr6_080096025	Chr6	80,096,025	Davis2	4.07	–	–	12.61	12.61	12.61	–	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr6_080096025	Chr6	80,096,025	Davis3	4.07	–	–	14.05	14.05	14.05	–	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr6_080096025	Chr6	80,096,025	ST2_GENO	4.12	–	–	0.07	0.07	0.07	–	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr7_006068288	Chr7	6,068,288	Davis4	4.13	–	–	–	–	–	-15.01	–	

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation	
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom		
	Chr8_014266142	Chr8	14,266,142	Davis1	4.09	–	–	23.06	23.06	23.06	–	–	
	Chr8_019911962	Chr8	19,911,962	ST2_GENO	4.18	–	–	–	–	–	-0.07	–	
	Chr8_061587300	Chr8	61,587,300	Davis3	3.39	–	–	–	–	-9.94	–	–	
	Chr8_061587300	Chr8	61,587,300	Davis4	3.39	–	–	–	–	-10.84	–	–	
	Chr8_084174416	Chr8	84,174,416	ST2_GENO	4.32	–	–	–	–	–	0.06	–	protein FATTY ACID EXPORT 2, chloroplastic-like
	Chr8_085866069	Chr8	85,866,069	Davis	4.45	–	–	–	–	–	–	–	EEIG1/EHBP1 protein amino-terminal domain protein
Persistence	Chr1_040541523	Chr1	40,541,523	Davis2	4.14	–	–	–	–	–	1.26	–	
	Chr1_073381901	Chr1	73,381,901	ST2_GENO	4.32	–	–	–	–	–	0.50	–	CRM-domain containing factor CFM3A, chloroplastic/mitochondrial isoform X2
	Chr2_007034650	Chr2	7,034,650	Davis5	4.12	–	–	–	–	–	-0.68	–	Hexosyltransferase (EC 2.4.1.-); PMR5N domain-containing protein; Mg-protoporphyrin IX

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom	
121												chelatase (EC 6.6.1.1); LRR receptor-like kinase
	Chr2_007034650	Chr2	7,034,650	ST2_GENO	4.18	–	–	–	–	–	-0.24	Hexosyltransferase (EC 2.4.1.-); PMR5N domain-containing protein; Mg-protoporphyrin IX chelatase (EC 6.6.1.1); LRR receptor-like kinase
	Chr2_032131303	Chr2	32,131,303	Westside1	4.29	–	–	–	–	–	0.66	–
	Chr2_042739112	Chr2	42,739,112	Davis3	4.29	–	–	-0.35	–	–	-0.35	–
	Chr2_042739112	Chr2	42,739,112	Davis4	4.28	–	–	-0.35	–	–	-0.35	–
	Chr2_042739112	Chr2	42,739,112	ST2_GENO	4.32	–	–	-0.13	–	–	-0.13	–
	Chr3_005267797	Chr3	5,267,797	ST2_GENO	4.32	–	–	–	–	–	0.75	F-box protein interaction domain protein; WPP domain-interacting protein 1
	Chr3_053349663	Chr3	53,349,663	Davis3	4.29	–	–	–	–	–	-0.98	–
	Chr3_053349663	Chr3	53,349,663	Davis4	4.28	–	–	–	–	–	-0.97	–

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation	
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom		
	Chr4_033456500	Chr4	33,456,500	ST2_GENO	4.49	—	—	-0.17	—	—	—	—	
	Chr4_044553513	Chr4	44,553,513	Westside	4.15	—	—	—	—	—	-0.37	glucomannan 4-beta-mannosyltransferase 2-like isoform X1	
	Chr5_051075912	Chr5	51,075,912	Davis2	4.28	—	—	—	—	—	0.98	—	
	Chr5_064929027	Chr5	64,929,027	Davis2	4.14	—	—	—	—	—	-0.83	—	
	Chr6_073984871	Chr6	73,984,871	ST2_GENO	4.32	—	—	—	—	—	-0.15	—	
	Chr8_039382141	Chr8	39,382,141	ST2_GENO	4.18	—	—	—	—	—	0.30	—	
	Chr8_067576904	Chr8	67,576,904	ST2_GENO	4.18	—	—	—	—	—	-0.79	—	
	Chr8_072762376	Chr8	72,762,376	Davis3	4.15	—	—	—	—	—	-1.06	Transportin-3; protein JINGUBANG-like isoform X1	
	Chr8_072762376	Chr8	72,762,376	Davis4	4.13	—	—	—	—	—	-1.15	Transportin-3; protein JINGUBANG-like isoform X1	

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom	
	Chr8_072762376	Chr8	72,762,376	ST2_GENO	4.18	–	–	–	–	–	-0.41	Transportin-3; protein JINGUBANG-like isoform X1
Plant height	Chr1_006491042	Chr1	6,491,042	Davis	4.09	–	–	–	0.00	0.00	–	Uncharacterized protein; F-box/LRR-repeat protein 15 isoform X3; PABS domain-containing protein;thermospermine synthase ACAULIS5;hypothetical protein
	Chr1_006491042	Chr1	6,491,042	Davis3	4.09	–	–	–	7.18	7.18	–	F-box/LRR-repeat protein 15 isoform X3; PABS domain-containing protein;thermospermine synthase ACAULIS5;hypothetical protein
	Chr1_009082943	Chr1	9,082,943	Westside	4.45	–	–	–	–	–	–	–
	Chr1_009082943	Chr1	9,082,943	Westside2	4.45	–	–	–	–	–	–	–
	Chr1_056974244	Chr1	56,974,244	Davis4	4.28	–	–	–	–	–	5.36	–
	Chr1_065651962	Chr1	65,651,962	Westside2	4.09	–	–	–	–	7.05	–	–

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation	
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom		
	Chr2_003485139	Chr2	3,485,139	Prosser	4.29	–	–	–	–	–	4.26	–	
	Chr2_012736370	Chr2	12,736,370	Westside	3.43	–	–	–	–	-5.31	–	–	
	Chr2_012736370	Chr2	12,736,370	Westside2	3.43	–	–	–	–	-6.39	–	–	
	Chr3_090153084	Chr3	90,153,084	ST2_GxE	4.18	–	–	–	–	–	-1.96		omega-hydroxypalmitate O-feruloyl transferase-like
	Chr5_025333490	Chr5	25,333,490	Prosser1	4.29	–	–	-6.30	–	–	-6.27		TRAF-type domain-containing protein
	Chr5_051075912	Chr5	51,075,912	Davis4	4.15	–	–	–	–	–	5.14	–	
	Chr6_050598088	Chr6	50,598,088	Prosser	4.15	–	–	3.09	–	–	3.43		Uncharacterized protein LOC115960095
	Chr6_050598088	Chr6	50,598,088	Prosser1	4.15	–	–	–	–	–	3.63		Uncharacterized protein LOC115960095
	Chr6_076896319	Chr6	76,896,319	Prosser1	4.15	–	–	–	–	–	2.91	–	
	Chr6_080096025	Chr6	80,096,025	Davis	4.09	–	–	–	–	0.00	–		TIR-NBS-LRR RCT1 resistance protein (Fragment)

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom	
Regrowth score	Chr6_080096025	Chr6	80,096,025	Davis3	4.09	–	–	10.59	10.59	10.59	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr6_080096025	Chr6	80,096,025	ST2_GxE	4.12	–	–	–	2.69	2.69	–	TIR-NBS-LRR RCT1 resistance protein (Fragment)
	Chr7_077351871	Chr7	77,351,871	Prosser1	4.29	–	–	–	–	–	-3.87	NBS resistance protein (Fragment)
	Chr1_040541523	Chr1	40,541,523	ST2_GENO	4.18	–	–	0.73	–	–	0.87	–
	Chr1_060661578	Chr1	60,661,578	Davis	4.09	–	–	–	–	0.65	–	Sorting nexin Vps5-like
	Chr1_060661578	Chr1	60,661,578	Davis3	4.09	–	–	–	1.36	1.36	–	Sorting nexin Vps5-like
	Chr2_001387962	Chr2	1,387,962	Prosser1	4.15	–	–	–	–	–	-0.57	–
	Chr2_003485139	Chr2	3,485,139	Prosser1	4.29	–	–	0.42	–	–	0.62	–
	Chr2_069084772	Chr2	69,084,772	Prosser1	3.43	–	–	–	–	-1.03	–	–
	Chr3_053349663	Chr3	53,349,663	ST2_GENO	4.32	–	–	–	–	–	-0.42	–

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom	
	Chr3_087196712	Chr3	87,196,712	ST2_GENO	4.32	–	–	–	–	–	-0.18	Superoxide dismutase (EC 1.15.1.1)
	Chr3_090153084	Chr3	90,153,084	Davis	4.15	–	–	–	–	–	-0.47	omega-hydroxypalmitate O-feruloyl transferase-like
	Chr3_090153084	Chr3	90,153,084	Davis3	4.15	–	–	–	–	–	-0.78	omega-hydroxypalmitate O-feruloyl transferase-like
	Chr3_090153084	Chr3	90,153,084	Davis4	4.15	–	–	–	–	–	-0.68	omega-hydroxypalmitate O-feruloyl transferase-like
	Chr3_090153084	Chr3	90,153,084	ST2_GENO	4.18	–	–	-0.26	–	–	-0.31	omega-hydroxypalmitate O-feruloyl transferase-like
	Chr3_090153084	Chr3	90,153,084	Prosser1	4.15	–	–	–	–	–	-0.40	omega-hydroxypalmitate O-feruloyl transferase-like
	Chr4_017342544	Chr4	17,342,544	Westside	4.09	–	–	0.80	0.80	0.80	–	–
	Chr4_017342544	Chr4	17,342,544	Westside2	4.09	–	–	1.28	1.28	1.28	–	–
	Chr4_021808044	Chr4	21,808,044	ST2_GENO	4.49	–	–	-0.15	–	–	–	–
	Chr4_044553513	Chr4	44,553,513	Davis	4.15	–	–	–	–	–	-0.61	glucomannan 4-beta-mannosyltransferase 2-like isoform X1

Trait	Marker	Chr	Position	Environment	Threshold	Effect size by genetic model						Protein annotation
						General	Diplo-gen.	Additive	Diplo-add.	1-Dom	2-Dom	
	Chr4_044553513	Chr4	44,553,513	Davis3	4.15	–	–	–	–	–	-1.13	glucomannan 4-beta-mannosyltransferase 2-like isoform X1
	Chr4_080360679	Chr4	80,360,679	ST2_GENO	4.18	–	–	–	–	–	0.70	–
	Chr6_030140972	Chr6	30,140,972	ST2_GENO	4.12	–	–	–	-0.77	-0.77	–	–
	Chr7_010103033	Chr7	10,103,033	Davis	4.15	–	–	–	–	–	-0.78	Uncharacterized protein
	Chr7_010103033	Chr7	10,103,033	Davis4	4.15	–	–	–	–	–	-1.22	Uncharacterized protein
	Chr8_005343637	Chr8	5,343,637	ST2_GENO	4.32	–	–	–	–	–	0.20	Disease resistance protein (TIR-NBS-LRR class)
	Chr8_023936224	Chr8	23,936,224	Westside	4.29	–	–	–	–	–	-0.27	–
	Chr8_023936224	Chr8	23,936,224	Westside2	4.29	–	–	–	–	–	-0.44	–

Effect sizes (additive g plant⁻¹ for dry weight; plants plot⁻¹ for persistence; cm plant⁻¹ for plant height; score units for regrowth score) are shown only for models in which the marker exceeded the Bonferroni-corrected threshold indicates the marker was not significant under that model. Chr, chromosome; Diplo-gen., diploid-general; Diplo-add., diploid-additive; 1-Dom, 1-dominance; 2-Dom, 2-dominance. Trait groups are separated by horizontal rules; alternating row shading aids readability.

Chapter 4

Genetic Variability Among Seed Lots of the Alfalfa Cultivar ‘CUF 101’

Kreingrai Nonkum, Dániel Pap, and E. Charles Brummer

Abstract

Plant identification is an essential process for maintaining the genetic integrity and commercial value of plant cultivars. Cultivar CUF 101, released in 1976, is a non-dormant alfalfa cultivar with high forage yield and strong pest and disease resistance. Since that time, CUF 101 has remained a widely grown cultivar, but concerns about its genetic integrity have been raised by seed suppliers and others. Experiments were conducted with the aim to identify putative cultivars all named CUF 101 by sequencing pooled DNA from 30–50 seedlings per entry across 84 seed lots. These included CUF 101 entries of varying certification classes. We also included check cultivars spanning a range of fall dormancy groups. A total of 1,588 SNPs were used to assess population structure with principal component analysis (PCA) and discriminant analysis of principal components (DAPC). PCA clearly distinguished CUF 101 from check cultivars, with the first principal component capturing the fall dormancy gradient. DAPC of CUF 101 entries revealed three distinct genetic clusters, with a subset of certified seed lots deviating substantially from the core CUF 101 genetic profile, suggesting that genetic contamination or admixture may have occurred over time. Variation in genetic purity was associated with seed company source and certification year, with 2016–17 certified lots showing greater genetic dispersion than 2014–15 lots.

1. Introduction

Alfalfa is one of the world's oldest domesticated crops and the third most important crop in the US (Putnam & Meccage, 2022). The alfalfa cultivar 'CUF 101' was developed by a collaboration between University of California, Davis, the US Department of Agriculture, and California farmers and was released in 1976 and was the first cultivar with resistance to the blue alfalfa aphid (Lehman et al., 1977; Lehman et al., 1983). CUF 101 is a non-dormant cultivar with fall dormancy score of 9 (Teuber et al., 1998) and is widely grown due to its adaptation to a wide range of growing conditions from the California Central Valley and low deserts to regions worldwide with similar environmental conditions. The consistent performance of CUF 101 and the relatively low cost of seed has led to its continued production and use by farmers. Initial breeder seed inventory was exhausted many years ago, and multiple foundation seed lots have been reproduced from previous foundation seed, from which certified seed is produced by multiple seed growers in the San Joaquin and Imperial Valleys of California. Through these generations, the genetic constitution of CUF 101 may have changed due to random genetic drift and selection. Contamination of seed lots may have occurred through outcrossing or mixing of non CUF 101 genotypes into the CUF 101 populations. Morphological identification is unlikely to clearly distinguish CUF 101 from other non-dormant alfalfa cultivars, but a genetic marker analysis could assist in identification of distinct populations.

The development of improved DNA genotyping methods based on sequencing (e.g., Elshire et al., 2011) enable rapid and accurate genetic analysis of plants and plant populations. A targeted genotyping method has been applied to alfalfa (Zhao et al., 2023) using the DArTag platform developed by Diversity Arrays Technology (Canberra, Australia, <https://www.diversityarrays.com>) as part of the DArTseq technology (Blyton et al., 2023). This

technique uses a previously identified pool of 3,000 alfalfa oligonucleotide sequences to capture the same set of genomic regions in all individuals being assayed, amplify their copy numbers, and bar-code them for sequencing (Krishnakumar et al., 2008; Zhao et al., 2023). Following purification and quantification, the products are sequenced using the Illumina platform to an average depth of approximately 10-100× per marker per sample (Hardigan et al., 2023). The resulting data are processed through DArT's pipeline for demultiplexing and variant detection. The alfalfa 3,000-marker (3K) DArTag panel has been evaluated in multiple autopolyploid alfalfa populations (Zhao et al., 2023). This method can identify allele dosage of individual plants or estimate allele frequencies of pooled genotypes. To make such data more accessible, Breeding Insight Genomics R Shiny application, BIGapp is a user-friendly, web-based platform that provides accessible genomic and bioinformatics tools designed for diploid and polyploid data (Sandercock et al., 2025).

The objective of this study was to assess the genetic relatedness of various generations and seed lots of CUF 101 to determine both their variability and their distinctiveness from other alfalfa cultivars.

2. Materials and Methods

2.1 Plant materials

In this experiment, 84 unique entries were included, consisting of 66 entries identified as CUF 101, along with 18 check cultivars and germplasms (Table 1).

Table 1. Seed entries and alfalfa cultivar groups evaluated in this experiment

Entry	Number of entries	Comments
CUF 101 Foundation	7	Seed lots representing different foundation seed generations maintained by the UC Davis Foundation Seed Program
CUF 101 Certified	20 (2014-2015) 23 (2016-2017)	Seed lots collected in 2014-2015 and 2016-2017 from seed growers in the Imperial Valley, CA by the California Crop Improvement Association during certification inspections.
CUF 101 Other	16	Seed lots labeled ‘CUF 101’ in the UC Davis seed storage room but for which little or no information is available on certification status or origin (although most probably derived from certified lots over the past 30 years)
Dormant & semi-dormant alfalfa cultivars with fall dormancy <6	9	Alfagraze, AmeriStand 403T, AmeriStand 518NT, HiGest 360, Gunner, Powell, Ranger, Vernal, WL 365HQ
Nondormant alfalfa cultivars with fall dormancy >6	9	AmeriStand 901TS, Highline, Magna 715, Magna 801FQ, Magna 995, UC Impalo, UC5796, UF_ALF_Pers2015, WL 656HQ

For each entry, seeds were fumigated with chlorine gas, prepared by mixing 100 mL of bleach with 3 mL of HCl, for 12 hours and then germinated in Petri dishes for seven days. 30-50 young seedlings were collected for each entry, pooled, and freeze dried for DNA extraction, which was performed using the DNeasy Plant Kit with a silica plate method (Qiagen, 69181) modified from CTAB method (Doyle 1991). The resulting DNA samples were screened for quality and adjusted to at least 50 ng/μl for further analysis.

2.2 DArTag genotyping

The DNA samples were submitted to DArT for DNA library preparation, target capture, and sequencing using the 3,000 SNP DArTag genotyping platform for alfalfa developed by the USDA Breeding Insight program (Zhao et al., 2023). Sequencing reads were subsequently processed by DArT to generate a Missing Allele Discovery Count (MADC) file. The MADC file contains read counts for reference and alternate alleles at the target SNP for each locus. In addition, off-target SNPs flanking the 3,000 target genomic regions are identified and the entire unique sequence of SNP alleles is referred to as a microhaplotype (Zhao et al., 2023).

2.3 Data filtering

The MADC file was filtered to retain loci if they were present in at least 80% of the samples at a minimum read depth of 30. Samples with more than 15% missing data were removed from further analysis. The filtered MADC file, including the genomic position and observed read counts for each locus, was imported into the polyRAD v.2.0.0 R package (Clark et al., 2019), following the pipeline adapted from Medina et al. (2025) to create a RADObject. We evaluated allele frequency as discrete classes by estimating the “genotype” of each pooled sample based on allele dosage with functions ‘addPCA’ and ‘IteratePopStruct’ with overdispersion parameter of 6. This polyRAD workflow enabled genotype calling while accommodating missing data. Genotype calls at the microhaplotype level were exported as a Variant Call Format (VCF) file using the ‘RADdata2VCF’ function with the option *asSNPs = FALSE*. In addition, genotype dosages of target SNPs were called following the microhaplotypes genotype calling but excluding RefMatch and AltMatch reads, retaining only Ref and Alt reads from the filtered MADC file.

The VCF file containing polyRAD genotypes was filtered using the “VCF Filtering” tab in BIGapp in R software (Sandercock et al., 2025). SNPs with a Minor allele frequency (MAF) below 0.05 were removed. From the 84 samples, 49 entries were sequenced in duplicate to assess marker assay repeatability, while the remaining 35 entries were represented by a single sample. Allele frequencies of replicate samples were compared by calculating the correlation between their genotypic profiles. The final dataset comprised 1,588 SNPs across 84 samples (prior to replicate averaging), representing 84 unique entries.

2.4 Genetic diversity

Summary metrics for filtered target SNP VCF files were evaluated using the “Genomic Diversity” tab in BIGapp. For each SNP, MAF was for each sample.

2.5 Population structure

Population structure analysis was performed using Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) (Jombart et al., 2010). PCA was conducted for target SNPs markers using ‘addPCA’ in polyRAD V 2.0.0 (Clark et al., 2019). DAPC was applied only to the target SNPs because the method requires biallelic markers. This analysis was carried out in the BIGapp “DAPC” tab, which implements the *adegenet* R package V 2.1.7 (Jombart, 2008; Jombart and Ahmed, 2011). For DAPC, marker matrices, marker positions, and entry IDs were converted into a *genlight* object. The BIGapp ‘DAPC’ tab first determined the optimum number of clusters (Best K) followed by the DAPC analysis using selected K. Both the PCA and DAPC scatterplots were generated using the *ggplot2* package in R.

3. Results

3.1 Marker information and reproducibility

After SNP filtering, 1,588 SNPs were retained across eight chromosomes of the *XinJiangDaYe* alfalfa reference genome, with SNP density ranging from 1.48 SNPs/Mb on chr6 (the lowest) to 3.36 SNPs/Mb on chr1 (the highest) (Table 2).

Table 2. Summary of the distribution of the 1,588 SNPs filtered target SNPs across the *XinJiangDaYe* alfalfa reference genome.

Chromosome	Length (bp)	SNPs	SNP Density (SNP/Mb)
chr1	82459472	277	3.36
chr2	76462061	252	3.30
chr3	93149498	290	3.11
chr4	90245664	281	3.11
chr5	81211777	260	3.20
chr6	80303593	119	1.48
chr7	88407277	243	2.74
chr8	87242343	287	3.29

To assess the reproducibility of allele frequency estimates, we sequenced two biological replicates for 49 of the 84 unique entries. Pearson correlation coefficients between replicate allele frequencies across 1,588 loci ranged from 0.92 to 0.99, with a mean of 0.96 (Figure 1), indicating

consistently high concordance between replicates. Given this strong agreement, allele frequencies from the two biological replicates were averaged for all subsequent analyses.

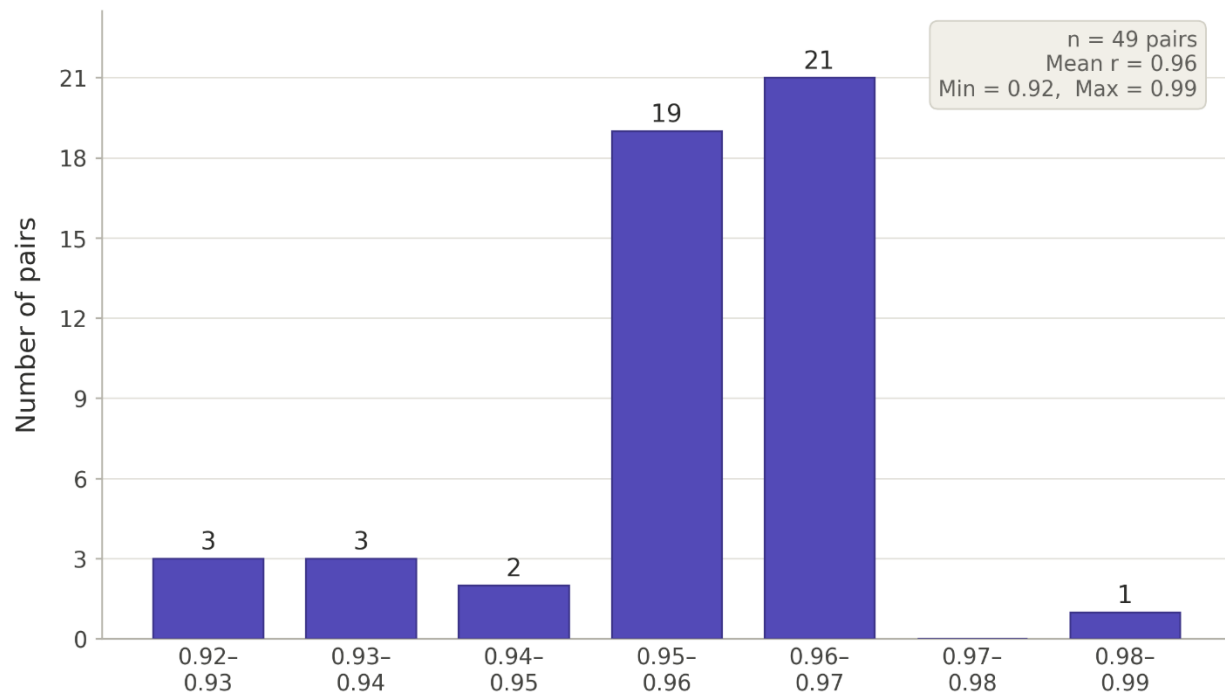


Figure 1. Allele frequency reproducibility in CUF 101 alfalfa (*Medicago sativa*) as a basis for population structure analysis

3.2 Population Structure

A principal component analysis (PCA) was performed across all entries to evaluate the genetic relationships among CUF 101 seed lots and other alfalfa cultivars. The first and second principal components explained 19.53% and 6.66% of the total genetic variance, respectively (Figure 2a). To more clearly see the relationships, we highlight the check cultivars with the CUF 101 Foundation seed lots (green ellipse) (Figure 2b). In this figure, 95% confidence ellipse of the three groups showed that cultivars were differentiated by dormancy class on PC1, with more dormant

cultivars to the right and less dormant to the left, and that Foundation CUF 101 seed lots formed a relatively tight sub-cluster within the broader nondormant genetic space. The general relationships among dormancy classes are expected, supporting the robustness of this analysis. The foundation seed lots differed somewhat from one another, but the reasonably tight cluster suggests that the population has not changed dramatically over generations. (Figure 2c) Two check cultivars, both also developed by the University of California, Cibola and Highline, cluster overlapped within the CUF 101 Foundation group but all other nondormant cultivars evaluated fall outside.

The certified (and other) CUF 101 seed lots show a relatively tight cluster in the upper-left region of the PCA plot (Figure 2c), consistent with the CUF 101 genetic space defined by Foundation seed lots. However, compared to Foundation lots, certified CUF entries showed a substantially broader distribution and a subset of certified CUF entries was positioned farther from the core cluster, with some overlapping with several non-UC non-dormant check cultivars, suggesting that genetic contamination or admixture may have occurred in those specific seed lots. The Other CUF entries generally co-located with known certified CUF entries.

To assess whether these outlier CUF 101 samples were truly distinct from other CUF 101 seed lots, we conducted a DAPC analysis of just the CUF 101 samples alone (Figure 3). This analysis suggested the presence of three clusters. Clusters 1 and 2 contained most samples, with Foundation samples largely in Cluster 1. However, two Foundation samples were in Cluster 2, which largely consisted of certified and other seed lots. These two clusters are nearly identical based on the first discriminant axis (LD1), which accounted for 91% of the variation and were only differentiated in the second dimension.

The outlier CUF 101 seed lots form a distinct Cluster 3, with eight certified lots and one other lot of unknown origin (Figure 3). These eight certified seed lots primarily derived from 2016-17 seed

lots with two from 2014-2015. Five of these lots derived from seed grower A, two from seed grower B, and one from seed grower E. Seed company A exhibited the widest dispersion, with several entries extending far along PC1 into genetic space more typical of non-CUF germplasm (Supplemental Figure 1). Entries from Seed Companies B, C, D, and E clustered more tightly in the central region of the PCA space, with Company C having the tightest clustering samples with the variation among seed lots was greater in 2016-17 than 2014-15 (Supplemental Figure 2).

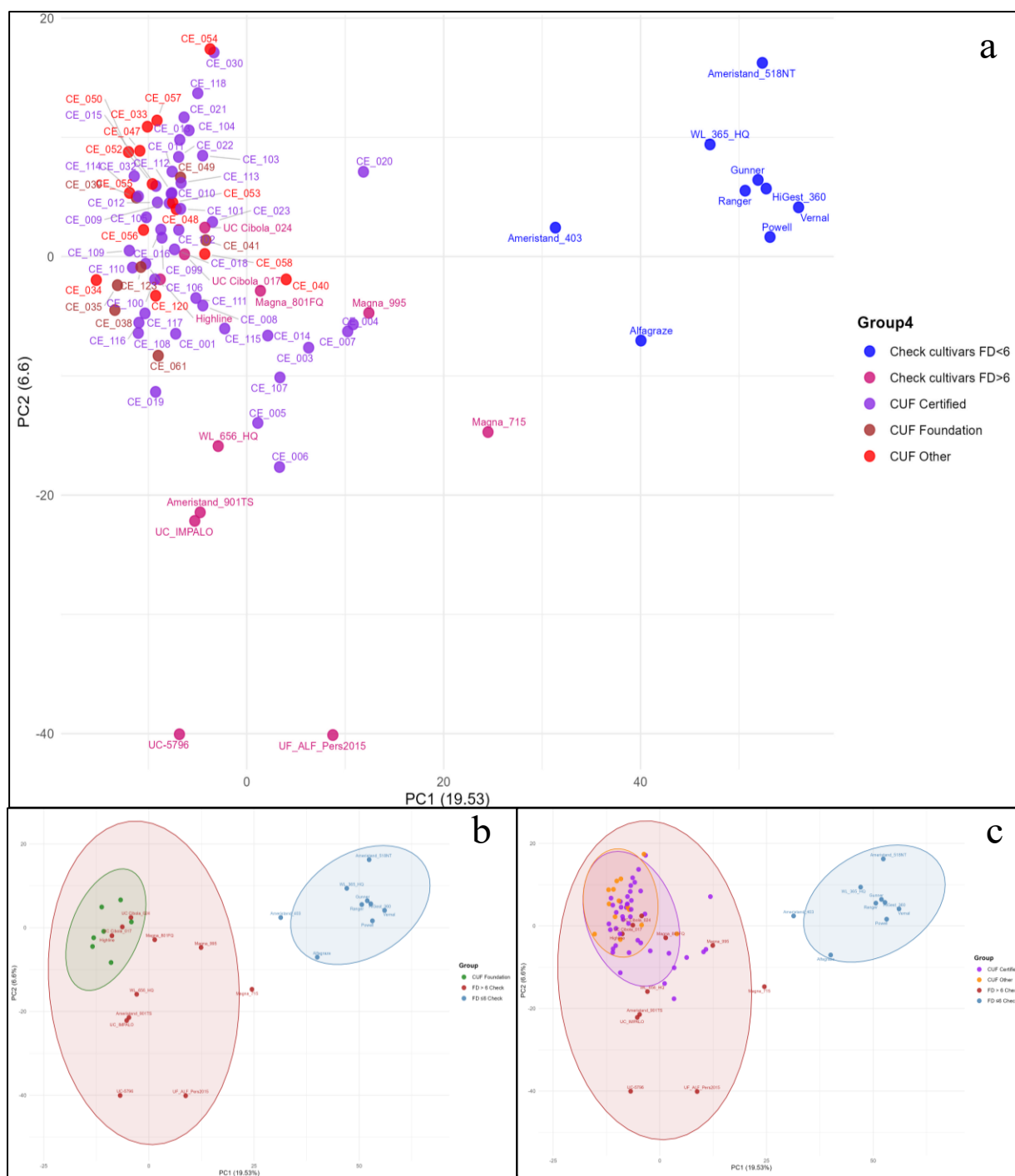


Figure 2. (a) The first two principal components from a PCA on all entries evaluated. (b) The same PCA but with only check cultivars and CUF 101 Foundation Seed lots with the year the foundation lot was produced. (c) The same PCA included checks and CUF certified and other seed lots.

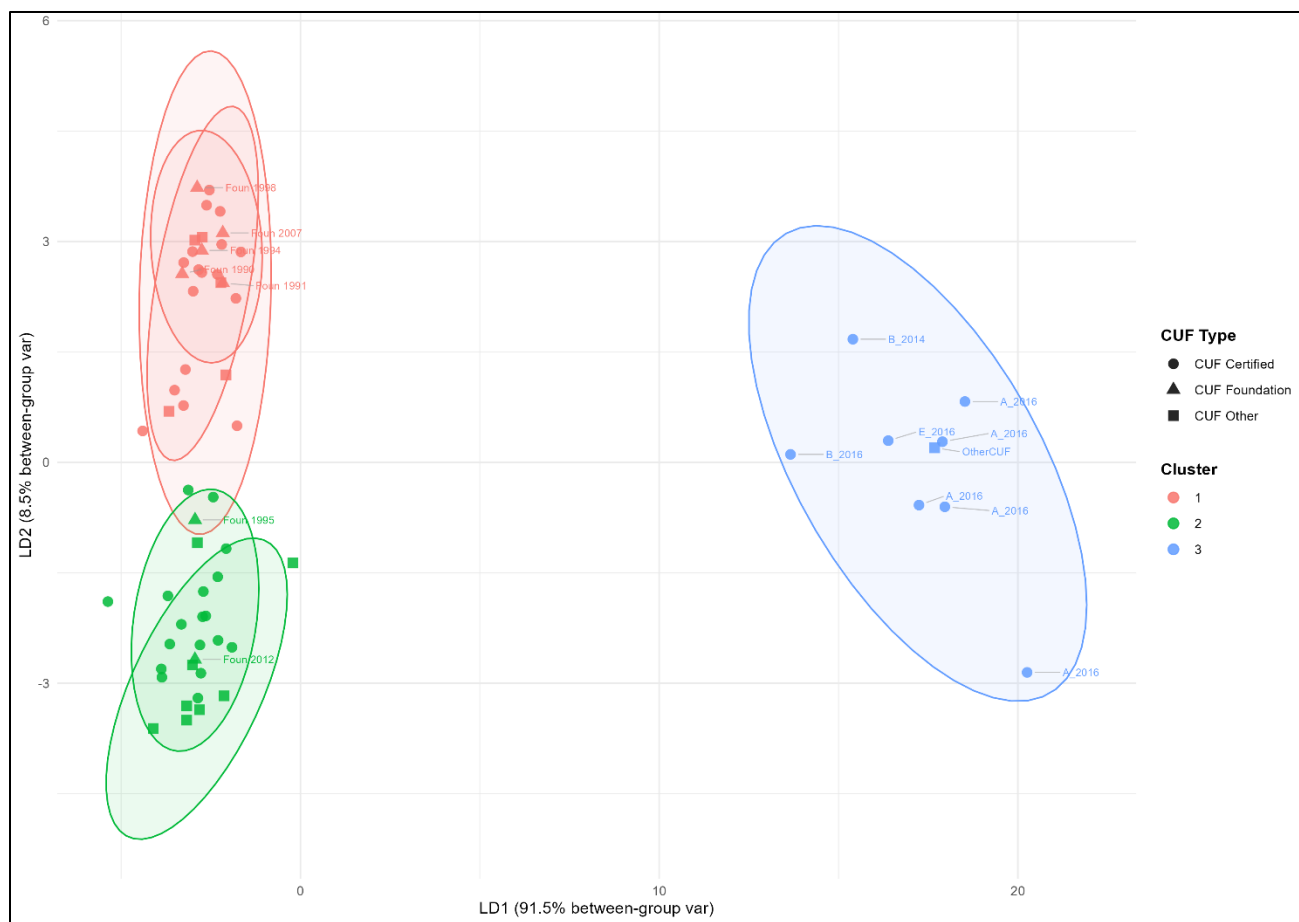


Figure 3. The first two loadings from a DAPC of CUF 101 seed lots using the most likely number of clusters (k) of 3, with three different types of CUF 101 seed indicated by symbols. Within clusters, multiple ellipses are drawn around each seed class.

4. Discussion

The objective of this study was to determine whether sequentially produced foundation seed lots of CUF 101 have changed genetically across generations and if certified seed lots vary genetically among seed companies and years. The relationship of CUF 101 seed lots to commercial cultivars across the fall dormancy spectrum was also assessed.

The reliability of the DarTag genotyping could be assessed in two ways. First, replicate samples showed a high correlation of allele frequencies, suggesting that the filtering criteria we used were sufficient to obtain reliable frequency estimates (Figure 1). Second, in the PCA analysis, cultivars separated broadly by fall dormancy with non-dormant cultivars ($FD > 6$) clustered on the negative end of PC1 alongside CUF 101, while dormant cultivars ($FD \leq 6$) clustered toward the positive end (Figure 2b). This result is consistent with Medina et al. (2025), who likewise identified fall dormancy as one of the strongest drivers of genetic differentiation among alfalfa populations genotyped with the DArTag platform, in both their DAPC and PCA analyses. These two results suggest that the relationship of CUF 101 seed lots should be reasonably robust.

Within the CUF 101 group, Foundation seed lots formed the tightest sub-cluster (Figure 2b). While foundation seed is expected to be true to type, the fact that multiple lots were sequentially produced which is one foundation seed lot was used to produce a “new” foundation seed lot, some genetic changes due to natural selection in the field, “migration,” and/or genetic drift would be expected. That these changes did not appear to be large is heartening. However, although the different lots cluster together, they are not identical. Whether this level of differentiation is acceptable is likely a matter of subjective judgement. Although the genetic markers we used here appear to be quite reliable, deeper sequencing coverage and more loci would likely provide a more fine-tuned understanding of the relationships among the seed lots.

The certified and other CUF entries showed a broader genetic spread indicating that considerable genetic variability accumulates during commercial seed multiplication. This is consistent with the open-pollinated nature of alfalfa, where repeated generations of seed production, even with the certified seed isolation protocols, can lead genetic divergence both from the foundation seed from which it originated and from other certified lots. The differences among seed lots of other cultivars should be examined to see whether the results here are typical of the variation seen across the alfalfa seed industry.

Several seed lots showed substantial differences from the bulk of CUF 101 seed lots; interestingly, these lots were also clustered, suggesting some similarity in their genetic constitution, while also recognizing that the cluster is quite broad. These accessions clustered close to several commercial cultivars, including Magna 801FQ, Magna 995, and WL 656HQ, well-known nondormant cultivars. The contamination appeared to be most common in seed sampled from Seed Grower A, but lots from two other seed growers were similarly genetically distinct from the bulk of CUF 101 seed lots. The results suggest that inadequate isolation, contaminated foundation seed prior to planting, or admixture events must have occurred. These findings suggest that periodic molecular marker testing across seed lots from multiple companies would be valuable for maintaining CUF 101 identity in the commercial seed supply. More recent certification cohorts exhibited greater genetic variability, with most of the outlier seed lots from the 2016–17 collection.

Although pooled sample methods show promise for genetic diversity analyses, pooled methods require high-quality sequencing data to minimize errors caused by mixed samples. Lynch et al., (2014) suggested that pooled samples sequencing should achieve $>100\times$ read coverage to improve allele frequency estimation and read discovery rates. The DArTag platform used in this study meets these requirements as it provides high read coverage ($\sim 100\times$ per marker) with low missing data rate

(Hardigan et al., 2023), making it well suited for pooled sample genetic analysis. Fracassetti et al. (2015) recommended pooling at least 25 individuals with high sequencing read depth ($> 100\times$) to achieve reliable SNP frequency estimates.

5. Conclusion

Principal component analysis clearly distinguished CUF 101 from a panel of dormant and non-dormant check cultivars. The CUF 101 entries resolved into three distinct genetic clusters, with one cluster of certified seed lots having potentially compromised genetic integrity. Variation in genetic purity was associated with both seed company sources and certification years, highlighting the value of routine molecular monitoring in the commercial seed supply. These results support the broader application of targeted genotyping approaches and user-friendly bioinformatics platforms for cultivar identification, seed lot certification, and quality assurance in alfalfa and other outcrossing polyploid species.

6. Acknowledgement

This research was funded by Brummer's lab at University of California, Davis, UC Davis Seed foundation. And Kreingkrai Nonkum was funded by The Royal Thai Scholarship from Thailand.

7. References

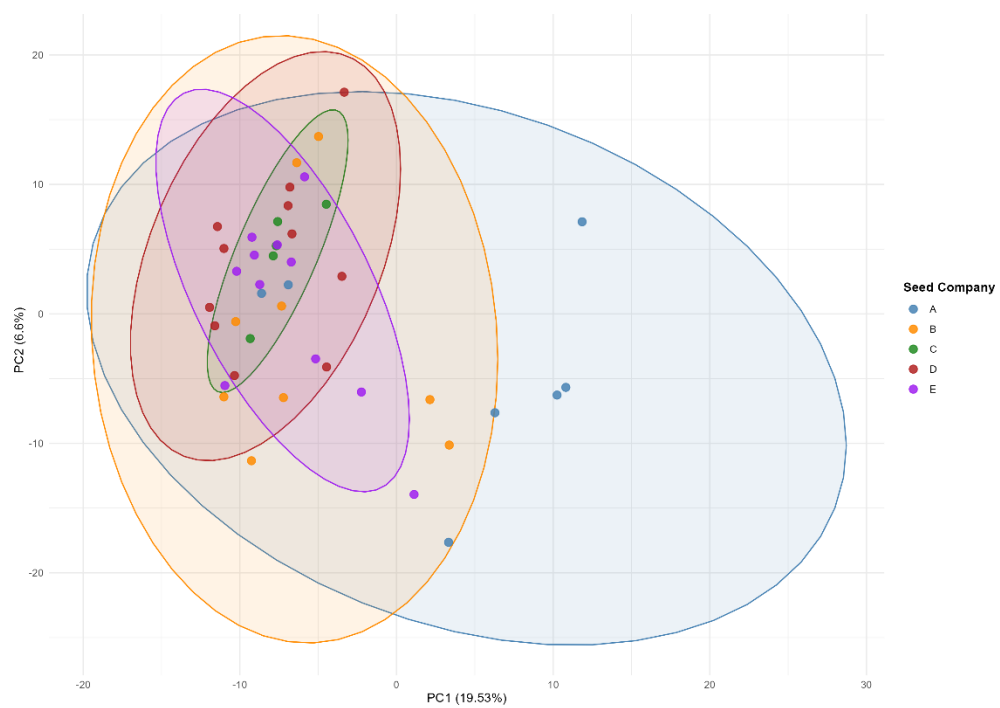
Bastien M, Boudhrioua C, Fortin G, Belzile F. (2018). Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. *Genome*. 2018 Jun;61(6):449-456. doi: 10.1139/gen-2017-0236. Epub 2018 Apr 24. PMID: 29688035.

- Blyton, M. D., Brice, K. L., Heller-Uszynska, K., Pascoe, J., Jaccoud, D., Leigh, K. A., & Moore, B. D. (2023). A new genetic method for diet determination from faeces that provides species level resolution in the koala. *bioRxiv*, 2023.02. 12.528172.
- Clark, L. V., Lipka, A. E. & Sacks, E. J. (2019). polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3 Genes Genomes Genet.* **9**, 663–673
- Correa Abondano, Miguel, Jessica Alejandra Ospina, Peter Wenzl, and Monica Carvajal-Yepes. (2024). "Sampling strategies for genotyping common bean (*Phaseolus vulgaris* L.) Genebank accessions with DArTseq: a comparison of single plants, multiple plants, and DNA pools." *Frontiers in Plant Science* 15: 1338332.
- DArT. (2023). *Diversity Array Technologies*. <https://www.diversityarrays.com>
- Darrier B, Russell J, Milner SG, Hedley PE, Shaw PD, Macaulay M, Ramsay LD, Halpin C, Mascher M, Fleury DL, Langridge P, Stein N and Waugh R (2019). A Comparison of Mainstream Genotyping Platforms for the Evaluation and Use of Barley Genetic Resources. *Front. Plant Sci.* 10:544. doi: 10.3389/fpls.2019.00544
- Doyle, J. J. and J. L. Doyle, (1990) "Isolation of Plant DNA from Fresh Tissue," *Focus*, Vol. 12, No. 1, 1990, pp. 13-15.
- DNeasy Plant Handbook, (2006)., <https://dna.uga.edu/wp-content/uploads/sites/51/2013/12/DNeasy-Plant-Handbook-for-DNA-purification-Qiagen.pdf>
- Doyle, J. (1991). "DNA protocols for plants." In *Molecular techniques in taxonomy*, pp. 283-293. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991.

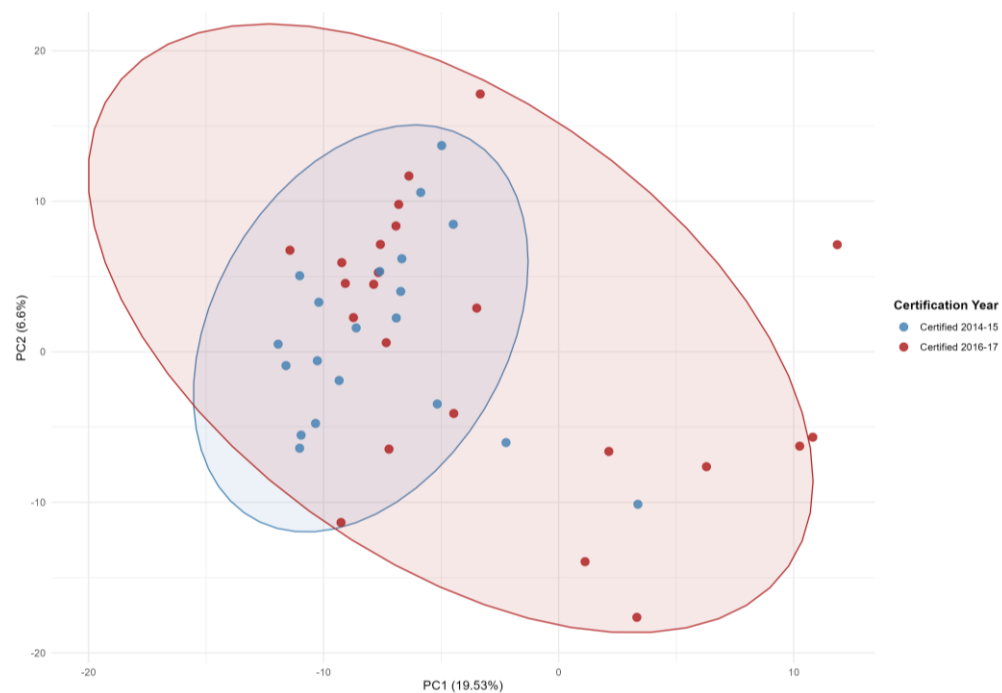
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, 6(5), e19379.
- Fracassetti M, Griffin PC, Willi Y (2015). Validation of Pooled Whole-Genome Re-Sequencing in *Arabidopsis lyrata*. PLOS ONE 10(10): e0140462. <https://doi.org/10.1371/journal.pone.0140462>
- Hardigan, M. A., Feldmann, M. J., Carling, J., Zhu, A., Kilian, A., Famula, R. A., Cole, G. S., & Knapp, S. J. (2023). A medium-density genotyping platform for cultivated strawberry using DArTag technology. *The Plant Genome*. <https://doi.org/10.1002/tpg2.20399>
- Jombart, T., Devillard, S. & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **11**, 94 <https://doi.org/10.1186/1471-2156-11-94>
- Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403–1405.
- Jombart, Thibaut, and Ismaïl Ahmed. (2011). "adegenet 1.3-1: new tools for the analysis of genome-wide SNP data." *Bioinformatics* 27.21: 3070-3071.
- Knaus BJ, Grünwald NJ (2016). “VcfR: an R package to manipulate and visualize VCF format data.” *BioRxiv*. <https://dx.doi.org/10.1101/041277>.
- Knaus BJ, Grünwald NJ (2017). “VCfR: a package to manipulate and visualize variant call format data in R.” *Molecular Ecology Resources*, 17(1), 44–53. ISSN 757, <https://dx.doi.org/10.1111/1755-0998.12549>.

- Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M., & Davis, R. (2008). A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proceedings of the National Academy of Sciences*, *105*(27), 9296–9301.
- Lehman, W. F., Nielson, M. W., Marble, V. L., & Stanford, E. H. (1983). Registration of Cuf 101 Alfalfa1 (Reg. No. 119). In *Crop Science: Vol. 23(2)* (Issue 2, p. crops1983.0011183X002300020054x).
<https://access.onlinelibrary.wiley.com/doi/abs/10.2135/cropsci1983.0011183X002300020054x>
- Lehman, W., Nielson, M., Marble, V., & Stanford, E. (1977). CUF 101, a new variety of alfalfa is resistant to the blue alfalfa aphid. *California Agriculture*, *31*(8), 4–5.
- Lynch M, Bost D, Wilson S, Maruki T, Harrison S. (2014). Population-genetic inference from pooled-sequencing data. *Genome Biol Evol.* 2014 Apr 30;6(5):1210-8. doi: 10.1093/gbe/evu085. PMID: 24787620; PMCID: PMC4040993
- Medina, C.A., Zhao, D., Lin, M. *et al.* (2025). Pre-breeding in alfalfa germplasm develops highly differentiated populations, as revealed by genome-wide microhaplotype markers. *Sci Rep* **15**, 1253 <https://doi.org/10.1038/s41598-024-84262-x>
- Meirmans, P. G. (2020). GENODIVE version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. *Mol. Ecol. Resour.* **20**, 1126–1131

- Putnam, D., & Meccage, E. (2022). Profitable alfalfa production sustains the environment. *Proceedings, 2022 World Alfalfa Congress*, 14–17.
- Sandercock, A. M., Peel, M. D., Taniguti, C. H., Chinchilla-Vargas, J., Chen, S., Sapkota, M., Lin, M., Zhao, D., Ackerman, A. J., Basnet, B. R., Beil, C. T., & Sheehan, M. J. (2025). BIGapp: A user-friendly genomic tool kit identified quantitative trait loci for creeping rootedness in alfalfa (*Medicago sativa* L.). *The Plant Genome*, 18, e70067. <https://doi.org/10.1002/tpg2.70067>
- Teuber, L. R., Taggard, K. L., Gibbs, L. K., McCaslin, M. H., Peterson, M. A., & Barnes, D. K. (1998). Fall dormancy. *Standard Tests to Characterize Alfalfa Cultivars*, A-1.
- Zhao, D., Mejia-Guerra, K. M., Mollinari, M., Samac, D., Irish, B., Heller-Uszynska, K., Beil, C. T., & Sheehan, M. J. (2023). A public mid-density genotyping platform for alfalfa (*Medicago sativa* L.). *Genetic Resources*, 4(8), 55–63.
- Zhao, D., Sandercock, A. M., Mejia-Guerra, M. K., Mollinari, M., Heller-Uszynska, K., Wadl, P. A., Webster, S. A., Beil, C. T., & Sheehan, M. J. (2024). A Public Mid-Density Genotyping Platform for Hexaploid Sweetpotato (*Ipomoea batatas* [L.] Lam). *Genes*, 15(8), 1047. <https://doi.org/10.3390/genes15081047>



Supplemental Figure 1. PCA of Certified CUF 101 samples labeled according to seed producer.



Supplemental Figure 2. PCA of Certified CUF 101 seed lots labeled according to year of production, 2014-15 vs 2016-17.

Concluding Remarks

This dissertation set out to evaluate whether a single mid-density genotyping platform could serve the diverse analytical needs of alfalfa breeding, and the evidence from three applications indicates that it can. First, the platform proved effective for characterizing germplasm diversity and population structure. Second, it was well suited to dissecting the polygenic basis of agronomic traits and predicting them via genomic prediction. Third, it could be used to verify cultivar identity and seed-lot integrity. Each application also exposed the boundaries of the current approach, chiefly the tension between the diversity that makes germplasm valuable and the marker density required to fully exploit it, together with the cost constraints that make pooled genotyping necessary in an obligate outcrossing crop.

The path forward is therefore one of integration rather than replacement: denser markers where resolution is limited, functional validation to convert associations into knowledge, deliberate broadening of the non-dormant genetic base, and the routine embedding of genomic tools across both the breeding pipeline and the seed supply chain. Pursued together, these directions offer a credible route toward renewed genetic gain in a crop whose productivity has been stagnant for a generation, and toward more resilient non-dormant cultivars suited to the warm, increasingly variable environments in which alfalfa is grown. The framework assembled in this work provides alfalfa breeders with a practical and adaptable foundation on which to build that future.