GWAS-assisted genomic prediction for forage nutritional quality traits in elite alfalfa (*Medicago sativa* L.) germplasm

Harpreet Kaur[1], Cesar Medina Culma[2], Christopher Pierce[1], Arshdeep Singh Gill[1], Jon Reich[3], Steve Damon[4], Long-Xi Yu[2], Ian Ray[1]

(1) Plant & Environmental Sciences Dept., New Mexico State University, NM, USA (2) United States Department of Agriculture-Agricultural Research Service, Prosser, WA, USA (3) Alforex Seeds, Woodland, CA, USA (4) Corteva Agriscience, West Salem, WI, USA

Introduction

- Alfalfa *(Medicago sativa L.)* is a highly valuable perennial forage legume. Most cultivars are autotetraploid (2n=4x=32).
- This crop is called "the Queen of Forages" because of its high nutritive value.
- Genome wide association study (GWAS) analysis and genomic prediction are promising plant breeding tools to improve nutritional quality traits in alfalfa.

Objective

- Identify genomic regions, and evaluate genomic prediction (GP) models, associated with nutritional quality traits in an elite New Mexico Genomic Study (NMGS) alfalfa population

Material and Methods

- **Plant material:** Three elite semi- & non-dormant populations random-mated 3 generations to create a New Mexico Genomic Study (NMGS) population. 215 NMGS maternal half-sib families were phenotyped.
- **Experimental Design:** Randomized Complete Block Design; repeating covariate check; three replicates.
- **Phenotyping**: In Las Cruces, NM, forage samples collected from the 3rd regrowth cycle in June 2018 and 2019, immediately prior to harvest.
- Forage samples assessed for 15 nutritional quality traits: ADF (Acid Detergent Fiber), aNDF (Adjusted Neutral Detergent Fiber), aNDFom (aNDF organic matter), Ash, Ca (Calcium), CP (Crude Protein), DNDF48 (Digestible NDF at 48 hours), IVTDMD48 (In Vitro True Dry Matter Digestibility at 48 hours), Lignin, K (Potassium), Mg (Magnesium), NDFD (NDF digestibility), P (Phosphorus), uNDF240 (Undigested NDF at 240 hours), uNDF240om (UNDF240 organic matter) using Near Infrared Reflectance Spectroscopy.
- **Statistical analysis:** Best linear unbiased estimates (BLUEs) for all traits in each year, using R based package **SpATS** (Spatial Analysis for field Trials using Splines)[1]
- **Genotyping by sequencing[2]:** Among maternal parents, 12,884 SNP markers identified using *Medicago truncatula* (Mtr4.0) reference genome assembly
- **Genome-Wide Association Study (GWAS)**: Marker-trait association determined using eight polyploid gene action models provided in the R package, GWASpoly[3]. False Discovery Rate (FDR) threshold of 0.05 was used to declare significant marker-trait associations (Score ~ 5).
- **Genomic prediction (GP):** All 12,884 SNP markers used to run various GP models including ridge regression best linear unbiased prediction (rrBLUP), genomic BLUP (GBLUP), Bayesian models, support vector machine (SVM)-linear, SVM-Gaussian, and random forest (RF) separately for each year.
- **GWAS-assisted GP**: A subset of SNP markers with score ($-\log_{10}$ of p-value) >1 were selected based on GWASpoly results to run the GP models. Also, weighted GBLUP (WGBLUP) conducted using scores from 6 GWASpoly gene action models as weights for all markers.
- **Cross validation (CV) scheme**: 10-fold CV repeated 500 times; 90% training and 10% validation data. Prediction accuracy is the Pearson correlation (r) between predicted and observed values in the validation data.

Acknowledgement

- Funding support - Alforex Seeds & New Mexico Agricultural Experiment Station
- Genotyping & phenotyping – Alforex Seeds, New Mexico State University, USDA-ARS (Prosser, Washington)

Results

Table 1. Summary statistics for 15 nutritional quality traits of 215 alfalfa maternal half-sib families in 2018 and 2019

Trait	2018			2019		
	Mean†	Range	CV (%)	Mean†	Range	CV (%)
ADF	31.81	29.28-35.24****	3.19	33.22	29.83-36.48****	3.89
aNDF	40.24	37.39-44.62****	3.02	38.57	34.88-42.83***	4.05
aNDFom	36.67	33.81-41.55****	3.75	37.59	33.44-42.56****	4.60
Ash	7.21	6.32-7.76****	3.26	12.22	11.03-13.17****	3.02
Ca	1.21	1.03-1.36****	3.66	1.35	1.15-1.50****	4.66
CP	22.07	19.4-23.75****	4.04	21.58	19.39-23.41***	4.00
DNDF48	11.84	10.84-12.85****	3.28	15.44	14.23-17.12****	3.37
IVTDMD48	79.53	75.4-82.07****	1.22	77.45	73.39-80.32****	1.66
Lignin	6.34	5.64-7.48****	4.04	6.37	5.47-7.50****	5.56
K	1.70	1.39-2.00****	7.48	2.55	2.19-2.84****	4.10
Mg	0.36	0.31-0.42****	3.78	0.37	0.32-0.43****	4.60
NDFD (% of NDF)	29.45	26.21-31.91****	2.98	40.08	36.3-43.09****	2.50
P	0.31	0.28-0.34****	2.76	0.33	0.31-0.35***	2.54
uNDF240	20.66	18.23-25.03****	5.04	22.99	20.17-27.61****	5.75
uNDF240om	18.31	15.81-22.69****	5.66	20.83	17.84-25.52****	6.52

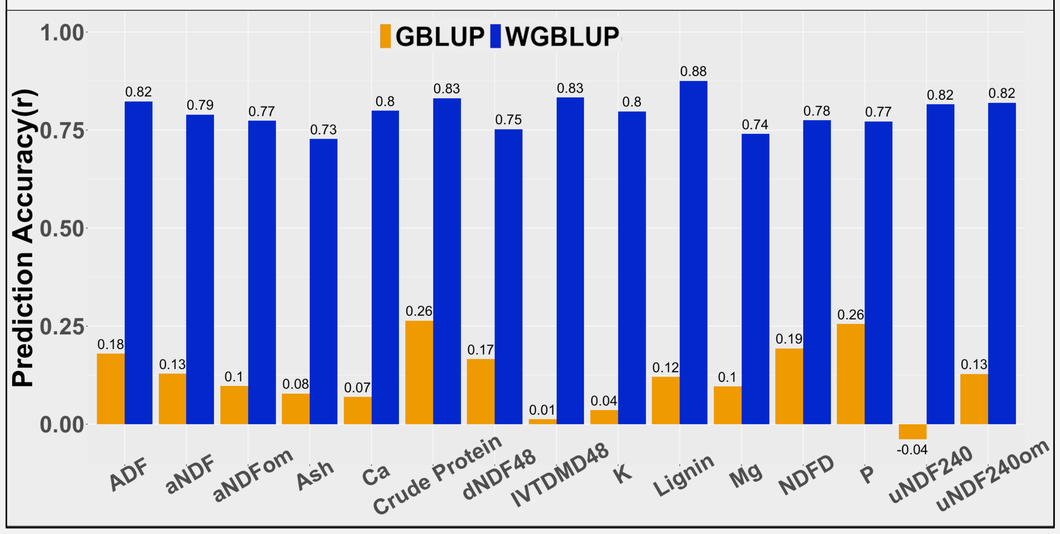
†Unit for means and ranges is % of dry matter; ***, **** significant at p<0.01 & p<0.001, respectively.

Table 2. Significant genome-wide associations between SNP markers and forage nutritional quality traits based on False Discovery Rate 0.05 threshold (Score ~ 5)

Year	Trait	Marker	Model‡	Score	Effect†
2018	NDFD	chr3_04866	SD-2	5.294	-0.81
	aNDF, CP	chr3_05000	DD-2, DD-2	4.600, 4.624	-1.395, 0.837
	CP, Lignin	chr4_06329	DD-2, DD-2	5.252, 5.167	0.625, -0.243
	CP, Lignin	chr4_06330	DD-2, DD-2	5.252, 5.167	0.625, -0.243
	IVTDMD48, Lignin	chr4_06395	DD-2, DD-2	4.702, 5.332	-0.631, 0.182
	Mg	chr5_08462	DD-2	5.082	-0.015
	CP	chr7_11284	DD-1	6.109	-0.758
2019	P	chr2_02793	DD-2	5.126	-0.005
	aNDFom	chr7_10742	DD-1	5.784	-1.583

† Eleven markers significant for only General and Diploidized General gene action models are not shown because their effects cannot be estimated. ‡ Models: **DD-1** = Duplex Dominant model with reference allele dominant over alternate allele, **SD-2** = Simplex Dominant model with alternate allele dominant over reference allele, **DD-2** = Duplex Dominant model with alternate allele dominant over reference allele.

Fig. 1. Average genomic prediction accuracies (r) for 15 nutritional quality traits in 2018 using all markers based on GBLUP (orange bars) and WGBLUP with marker weights from GWASpoly scores for an additive gene action model (blue bars).



Key Findings

- Significant differences observed in BLUEs for all 15 forage quality traits (Table 1).
- 25 SNP markers in 2018 and two in 2019 were significantly associated with 11 quality traits based on an FDR 0.05 threshold. Of these markers, 18 were detected only by General or Diploidized General gene action models with inestimable effects, and 9 were detected by models capable of estimating marker effects (Table 2).
- Multiple SNP markers were significantly associated with two or more forage quality traits (Table 2) e.g., marker chr3_05000 was associated with aNDF and CP.

Key Findings

- When SNP marker effects could be estimated, the direction of marker effects for different traits were in agreement with known relationships between forage quality traits. For instance, the chr3_05000 marker demonstrated a positive effect for CP and a negative effect for aNDF.
- When the FDR threshold was relaxed to Score > 1, 1,952 - 2,134 markers were identified for each trait in both years.
- Only 2 SNP markers were significantly associated with 2 traits in 2019; possibly due to heavy rain and high wind which caused shoot lodging a week prior to shoot sampling. However, many genomic regions were identified with Score > 1 SNP markers that showed consistent direction of forage quality effects in both years (data not shown).
- Such genome regions may be impacting forage quality and can be targeted for marker assisted selection for favorable regions and against unfavorable regions.
- Average prediction accuracy (r) for genomic prediction models using all the SNP markers were low to moderate ranging -0.03 to 0.31 for rrBLUP, -0.02 to 0.26 for GBLUP, -0.03 to 0.25 for SVM-Linear, -0.10 to 0.25 for SVM-Gaussian, -0.09 to 0.18 for RF models & -0.02 to 0.28 for Bayesian models (Fig. 1; only GBLUP results shown; orange bars) .
- GWAS-assisted GP using SNP marker subsets with score > 1 greatly enhanced the prediction accuracies for all the models, which ranged from 0.78 to 0.90 for rrBLUP, 0.86 to 0.94 for SVM-Linear, 0.76 to 0.83 for SVM-Gaussian, and 0.56 to 0.70 for RF. (data not shown)
- GWAS-assisted GP using all 12,884 SNP markers in a WGBLUP model, that employed scores from 6 GWASpoly gene action models as marker weights, also increased prediction accuracies (GP range: 0.57 to 0.88). Example: see Fig. 1 for WGBLUP with GWASpoly additive gene action model results (blue bars).
- After three cycles of random mating, significant linkage disequilibrium remains in the population. Significant markers most likely represent potential genome regions (not candidate genes) influencing forage quality traits.

Conclusion

- FDR correction is very stringent and may fail to detect many biologically relevant genome regions that impact alfalfa nutritional value.
- Relaxation of GWAS analysis score thresholds may facilitate identification of a subset of biologically relevant SNP markers which can be included in genomic selection models to improve prediction accuracies.
- WGBLUP models also enhance prediction accuracies, and with a large number of markers, they are computationally less intensive compared to machine-learning and Bayesian models.
- Independent validation of such genomic prediction models is needed.
- Contingent upon validation, selected SNP markers may be useful for developing elite alfalfa germplasm with enhanced nutritional value.

References

[1]Rodriguez-Alvarez, M.X. et. al., (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, 23, 52 - 71.

[2]Elshire, R. J. et al., (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

[3]Rosyara, U. R. et al., (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *The Plant Genome*, *9*(2), 1-10.