U.S. Alfalfa Farmer Research Initiative

Evaluating approaches to high-throughput phenotyping and genotyping for genomic selection in alfalfa

Contact information for PI and Co-PIs

PI: Kelly Robbins[1], Assistant Professor, (607) 255-8819, krr73@cornell.edu
Co-PI: Don Viands[1], Professor, (607) 255-3081, drv3@cornell.edu
Co-PI: Julie Hansen[1], Senior Research Associate, (607) 255-5043, jlh17@cornell.edu
Author: Nicholas Santantonio[2], Assistant Professor, (540) 231-5127 nsant@vt.edu
[1]233 Emerson Hall, Plant Breeding and Genetics, School of Integrated Plant Sciences, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY 14853.
[2]School of Plant and Environmental Sciences, College of Agriculture and Life Sciences, Virginia Tech, Blacksburg VA.

1 Abstract

Abstract: (Limit 200-300 words)

2 Introduction

[Need to acknowledge Noble and send to Maria Monteros! Double check MTA for compliance!]

Genetic gain in alfalfa has approached stagnation in the past few decades, limiting benefits to alfalfa farmers. Adoption of new breeding technologies has also lagged due to the complexity of the genetics, a high phenotypic burden and a paucity of public funds for a crop that is just one degree of separation too far from the consumer's mouth, and interest. Evaluation of breeding material requires multiple harvests per year for multiple years, limiting the size and number of field trials. The low heritability of forage yield also demands extensive replication, further limiting the number of breeding populations that can be evaluated. The ability to screen more material will lead to higher effective selection intensities, and increase the frequency of developing populations that outperform current varieties. This project aims to determine how affordable new technologies including high-throughput genotyping and phenotyping, can provide additional information to reduce the phenotypic burden while providing insight into how genetic variability of growth and development leads to differential forage yield.

High throughput phenotyping (HTP) technologies could drastically reduce the phenotypic burden in alfalfa by replacing a plot harvester with a unmanned aerial vehicle equipped with a multi-spectral camera for some harvests, locations, and/or replications. Quantitative genetic models can be built to accurately predict forage yields from spectral imaging, especially given that the harvested product is imaged directly. Images taken throughout the production years of a stand can also provide insight into genotype by environment interactions (G×E), in which varieties have differential growth responses under different conditions. Understanding the genetic signal in differential growth response will allow for identification of breeding targets and optimal population change for sets of predictable environmental conditions.

Inclusion of genome-wide markers can improve these types of prediction models by enabling related material to share information. These genomic prediction models can allow for reduced replication, sparse testing and even prediction of unobserved populations. Estimating realized genetic relationships in alfalfa is complicated by the fact that varieties are not genetically distinct individuals. As an obligate outcrosser, alfalfa is typically bred on a population level, where varieties are released as synthetics to avoid inbreeding and take advantage of population-level heterosis. This has limited implementation of marker-based selection because large numbers of individuals must be genotyped and inter-mated to avoid inbreeding in future generations. Single individuals are not representative of a variety as a whole, and genotyping many individuals from each variety is costly and restrictive.

As part of this study, we evaluated a new genotyping strategy for alfalfa, where DNA from many individuals is bulked in a given breeding population or variety for genotyping. Because much of quantitative genetics and selection theory hinges on population level parameters, the current machinery can be easily adapted to breeding on a population level. By borrowing ideas from population genetics, allele counts within each variety or breeding population, as opposed to allele counts within each individual, can be used to estimate genetic relationships between populations using pairwise Fst statistics (Weir and Hill 2002). This genotyping strategy should allow for prediction of additive effects for genetic gain, as well as dominance effects to exploit population level heterosis.

Development of an affordable, population-level genotyping method would need the ability to count alleles in a given sample, a task well suited for sequence-based methods. Whole-genome resequencing of the nine historic alfalfa germplasm sources (Barnes et al. 1977; Segovia-Lerma et al. 2004), as well as materials from the Cornell forage breeding project, were used as a proof of concept to determine the efficacy of a sequenced-based population-level genotyping while identifying sites most applicable to such a method. Whole-genome sequences will be made publicly available for greater use within the community. We are also currently collaborating with Breeding Insight to incorporate highly polymorphic regions identified in these materials that can be used in a wide array of North American alfalfa germplasm to help build an affordable genotyping platform.

In this report, we detail our findings for incorporating genome-wide population-level markers and high-throughput phenotyping to reduce phenotypic burden, estimate genotype specific growth curves, and how they are related to forage yield and quality.

Methodology

2.1 Plant materials

To help evaluate the efficacy of the proposed bulk genotyping method in a diverse background, remnant seed from a diallel study (Segovia-Lerma et al. 2004) was obtained from Ian Ray at New Mexico State University. Segovia Lerma et al. (2004) created a half diallel by crossing all possible pairs of the nine historic North American germplasm source populations, African, Chilean, Flemish, Indian, Ladak, M. *falcata*, M. *varia*, Peruvian, and Turkistan (Barnes et al. 1977). The resulting 36 hybrid populations along with the 9 parental populations were evaluated in the field in 1997 and 1998 near Las Cruces, NM in a replicated complete block design. Forage dry matter content data from this experiment for five harvests in each year was provided to us by Ian Ray at New Mexico State University (NMSU), as well as AFLP genotyping data from the nine parental populations (1544 AFLP markers; Segovia-Lerma et al. 2004).

Eight Cornell varieties and breeding populations were selected for sequence-based genotyping. These eight populations were established in a replicated variety trial along with seven commercial populations with 5 replicates in Geneva, NY in the spring of 2017. Remnant seed from the trial planting was obtained and used for sequence-based genotyping. Permission to genotype the remaining seven commercial varieties was not obtained at the time of project conception. Forage yield was measured using a plot flail harvester, and dry matter yield for each plot was calculated from fresh forage weight and dry matter content samples. Forage yield was collected for three cuts in 2018, 2019 and 2020. Only forage yields from cuts where aerial imaging was conducted are included in this report, which consisted of harvest two and three in 2019, and harvests 1 one and two in 2020. Quality samples from the second regrowth in 2019 and 2020 were harvested using standard practices, dried and ground. These samples were submitted to Dairy One for quantification of percent crude protein (CP) and percent neutral detergent fiber (NDF).

2.2 Population-level Genotyping

Whole genome resequencing of bulk samples from nine historic North American germplasm sources, five hybrid populations and eight Cornell varieties, was performed at Cornell University. One hundred seed from each population were germinated, and 25 seed with radicle extension of 1-5 mm were bulk homogenized in a single well for DNA extraction. DNA extraction and sequence library preparation was performed by the Bioinformatics Research Center at Cornell University. Sequencing of these 24 samples was performed on an Illumina NovaSeq 6000 with a S2 flowcell to produce approximate 1,000 Gbp of single paired-end 150 bp reads at Weill Cornell Medical to acheive approximately 50× coverage. Reads from each of the four bulk samples from each population were pooled for alignment and variant calling.

Sequences were aligned to the Bionano tetraploid alfalfa genome assembly 3a.2, kindly provided by the Noble Research Institute using the Burrows Wheeler alignment tool, bwa (Li and Durbin 2009). Only reads with a map quality over 20 (i.e. $p = 10^{-20}$ of mapping position being wrong) were kept to minimize alignment to multiple sites. Bi-allelic variant calls were performed using bcftools (Li 2011, SAMtools), and were further filtered based on a minimum and maximum read count of 20 and 125, respectively for all genotyped individuals. The minimum count allowed for reasonable estimation of allele frequencies, while the maximum reduced the probability of multiple alignment, given duplications not present in the reference genome. A final filter was imposed to remove sites with a global minor allele frequency greater than 0.05.

The hybrid populations were genotyped 'in silico' using the allele frequencies of the genotyped parental populations as

$$p_{a,k,ij} = \frac{1}{2}(p_{a,k,i} + p_{a,k,j}) \quad \forall\ i \neq j,$$

where $p_{a,k,i}$ and $p_{a,k,ij}$ is the i^{th} parent population allele frequency and the ij^{th} expected hybrid population allele frequency, respectively, for the a^{th} allele of the k^{th} marker. Correlation between the allele frequency estimates for the five hybrid populations that were sequenced, and their expected allele frequencies based on their parent populations were calculated to validate the 'in silico' approach.

Additive genetic relationships between populations were calculated from allele frequency estimates within each population. This was done using two methods. First, a simple covariance matrix between lines was calculated from the matrix of allele frequencies, $\mathbf{P}$, as $\mathbf{G} = n\mathbf{K}/tr(\mathbf{K})$, where $\mathbf{K} = (m-1)^{-1}\mathbf{P}'\mathbf{P}$ for m sites. Second, pairwise F_{st} statistics were used to estimate the additive genetic relationships between populations using the results from Weir and Hill (2002, equation 7).The latter is relative to an unknown average between population relatedness quantity.

2.3 Aerial phenotyping

Aerial phenotyping commenced on July 5th, 2019 during the in the second year of forage production shortly after the first cut on June 27th. A DJI Matrice 600 Pro unmanned aerial vehicle (UAV) equipped with a Micasense Rededge-MX multi-spectral camera was used for all flights. A flight plan was designed to obtain an 80% overlap in images collected at a flight speed of 2 m/s and an altitude of 20 m. Flights were conducted within 2 hours of solar noon on clear days when possible. A total of forty flights were conducted on average every 4.3 days across four harvests, with 10, 11, 12 and 7 flights for the cut 2 2019, cut 3 2019, cut 1 2020 and cut 2 2020, respectively. Four ground control points positioned at the four corners of the trial were measured with a Trimble RTK-GPS, which was used to geo-locate plots. Orthomosacis were constructed using Pix4D mapping software, and were subsequently uploaded into Imagebreed (www.imagebreed.org), a plot image database developed by our lab (Morales et al. 2020), for image processing and storage and vegetative index (VI) calculation at the plot level. All plot level image data has been made publicly available at www.imagebreed.org, while phenotypes and genotypes will be made publicly available at the time of publication, or by request.

Normalized difference vegetation indices (NDVI) were calculated from mean pixel values of near infrared (NIR) and Red bands of plot level images as

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

.

Normalized difference red edge indices (NDRE) were also constructed and analyzed, but were found to be less predictive than NDVI (results not shown).

2.4 Multivariate prediction model

A multivariate mixed model was used to estimate genetic correlation between end-use traits, and their genetic correlation to individual vegetative indices at different time points.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{g} + \mathbf{e} \tag{2}$$

where the phenotypic observations of $\mathbf{y} = [\mathbf{y}_1', \ldots, \mathbf{y}_t']'$ for t traits, and $\text{Var}(\mathbf{y}) = \mathbf{U} \otimes \mathbf{G} + \mathbf{R}$, where $\mathbf{U}$ and $\mathbf{R}$ are the unstructured trait and error covariance matrices to be estimated, and $\mathbf{G}$ is genetic the covariance between populations calculated from genetic markers.

When the number of traits to be considered is small, all traits can be included in a single multivariate mixed model to estimate the $k + \binom{k}{2}$ trait variance components [[and $k + \binom{k}{2}$ error variance components]]; however, as the number of traits increases the number of variance components to be estimated becomes intractable. For example, estimating the genetic correlation between ten vegetative index time points and forage yield requires estimation of 11 variance components and 55 covariance components, which requires large data sets. For this study, all traits were included in a single multivariate model if the number of traits was ≤ 4. When the number of desired comparisons was greater than 4, such as when comparing all vegetative index time points with forage yield, separate bivariate models were fit to estimate the genetic correlation in a pairwise fashion.

Bivariate models were also used for prediction of end-use traits, where a vegetative index was observed for all plots, but end-use traits were only observed for some plots.

2.5 Genomic prediction

Cross validation of genomic prediction in the diallel was used to compare a sequence-based approach for estimating genetic relationships to other strategies that cannot estimate allele frequencies (i.e. dominant markers). The dominant AFLP markers from Segovia-Lerma et al (2003) were used to construct hybrid population genotypes 'in silico' by summing the AFLP scores for each parental pair, and subsequently used to calculate a genetic covariance. A 'leave one family out' strategy, in which phenotypic observations from all hybrids formed from one of the nine parents, were removed and genetic values of those hybrids predicted.

[this is a little out of order, may need to rearrange]

To evaluate the potential of using vegetative indices to reduce the phenotypic burden of harvesting plots, two types prediction models were tested. The first model observed VIs for all replications in each trial, but did not observe plot forage yield for one to four replications. A bivariate linear mixed model was fit to estimate the genetic correlation between the VI and forage yield for a given harvest, and predict the genetic values of entries for forage yield with the missing replications. The second model observed both VI and forage yield for two to three harvests, but only VI for the remaining one or two harvests. The genetic correlation of VI with forage yield in observed harvests was then used to predict the genetic value of entries in harvests with only VI information. Genetic relatedness between lines was included for the genotyped set of eight entries, and not included for all fifteen entries. Correlations between genetic values estimated as genotypic means within harvest using all forage data and those predicted genetic values with missing forage yield data were used to assess predictability.

2.6 Growth Curves

Growing degree days (GDD) for each i^{th} flight were calculated from January 1^{st} of each year in Imagebreed as

Symbol	Degree	Legendre polynomial function
$\mathbf{l}_0$	0	1
$\mathbf{l}_1$	1	$\mathbf{x}$
$\mathbf{l}_2$	2	$\frac{1}{2}(3\mathbf{x}^2 - 1)$
$\mathbf{l}_3$	3	$\frac{1}{2}(5\mathbf{x}^3 - 3\mathbf{x})$

$$GDD_i = \sum_{d=1}^{D} \frac{max(C_d) - min(C_d)}{2} - 5°C \tag{3}$$

where D equals the Julian day of the flight, and C is the temperature in Celsius. Daily temperatures were sourced from the Geneva weather station on the site of the trial (USC00303184; Global Historical Climatology Network Daily, NOAA 2020). GDD were standardized to have a min and max of -1, and 1 for use as predictors for Legendre polynomials in the random regression model. The base temperature of $5°C$ is recommended for calculating growing degree days for field grown alfalfa (Sharratt, Sheaffer, and Baker 1989).

Genotype specific growth curves were fit using all vegetative indices within a regrowth cycle as well as the end-use phenotypes in the following random regression model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{g} + \mathbf{e} \tag{4}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_t \\ \mathbf{y}_p \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{1}_{(t+1)bn} & \mathbf{1}_{t+1} \otimes \mathbf{I}_b \otimes \mathbf{1}_g \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{l}_0 \; \mathbf{l}_1 \; \mathbf{l}_2 \; \mathbf{l}_3 \end{bmatrix} \tag{5}$$

and $\mathbf{y}_i$ are vectors of plot level vegetative indices for each $i \in 1, ..., t$ time points, and $\mathbf{y}_p$ is the vector of an end-use plot level phenotype. Vectors $\mathbf{l}_0$, $\mathbf{l}_1$, $\mathbf{l}_2$ and $\mathbf{l}_3$ are the evaluated functions, $f(x)$, of the first four legendre polynomials for x growing degree days, standardized between -1 and 1 when $\mathbf{y}$ is a vegetative index, and 0 otherwise (see Table ??). The vector $\mathbf{p} = 1$ when $\mathbf{y}$ is a phenotype, and 0 otherwise, and is used to estimate the genotypic effect of the end-use trait. The vectors, $\boldsymbol{\beta}$, of fixed block effects within each time point and end-use phenotype, u of Legendre function parameters for each genotype, and g of end-use trait estimates for each genotype are estimated using restricted maximum liklihood implemented in remlF90 (Misztal et al. 2002).

An important benefit of using a random regression model with covariance functions, is that the number of variance components that need estimated is limited $(l+1) + \binom{l+1}{2}$ where $l =$ to the degree of the legendre polynomial functions plus one for the

No permanent environmental effect beyond block time effects was fit due to the relatively small physical size of the trial and number of entries.

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}\\ \mathbf{g}\end{pmatrix} = \begin{bmatrix} \sigma^2_{c1} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{1g} \\ & \sigma^2_{c2} & \sigma_{23} & \sigma_{24} & \sigma_{2g} \\ & & \sigma^2_{c3} & \sigma_{34} & \sigma_{3g} \\ & \text{Symm.} & & \sigma^2_{c4} & \sigma_{4g} \\ & & & & \sigma^2_{g} \end{bmatrix} \otimes \mathbf{G} \tag{6}$$

where σ^2_{ci} and σ_{ij} are the variance and covariance of the Legendre polynomial coefficients, σ^2_g and σ_{gj} is the genetic variance of the end-use trait and the covariance between the end-use trait and the Legendre coefficients, respectively, and $\mathbf{G}$ is the genetic covariance of lines calculated from allele frequencies estimates.

These random regression models allow for information sharing across genetically related populations to estimate parameters of growth curve functions for each population, as well as their correlation with the end-use phenotype.

Objectives

The objective of this study was to evaluate the potential for combining a population-level genotyping approach with aerial imaging to model growth and development, reduce phenotypic burden and aid in genomic selection strategies.

Project Objectives:

1. Evaluate the efficacy of using a sequence-based population-level bulk genotyping approach to predict yield performance in diverse and elite germplasm.
2. Estimate the genetic correlations of multi-spectral indices with forage yield and quality using population-level genomic relationships.
3. Determine efficacy of phenotype reduction using spectral indices.
4. Fit population specific growth curves for each harvest using genomic relationships and spectral indices.

Project Results:

1. Pairwise F_{st} values serve as efficient estimates of genetic relatedness between populations.
2. Genetic correlations between forage yield and vegetative indices are high, especially toward beginning of regrowth period
3. Vegetative indices
4. Early growth tends to lead to higher forage yields, but with lower quality. Quality reduction likely related to maturity.

3 Results

3.1 Population-level genotyping

Estimates of hybrid population allele frequencies based on parental population frequencies were highly correlated with those observed in the five hybrid populations from the diallel that were included in the sequencing, ranging from 0.88 to 0.91. This suggests the bulk genotyping approach is effective at sampling the true allele frequencies within populations.

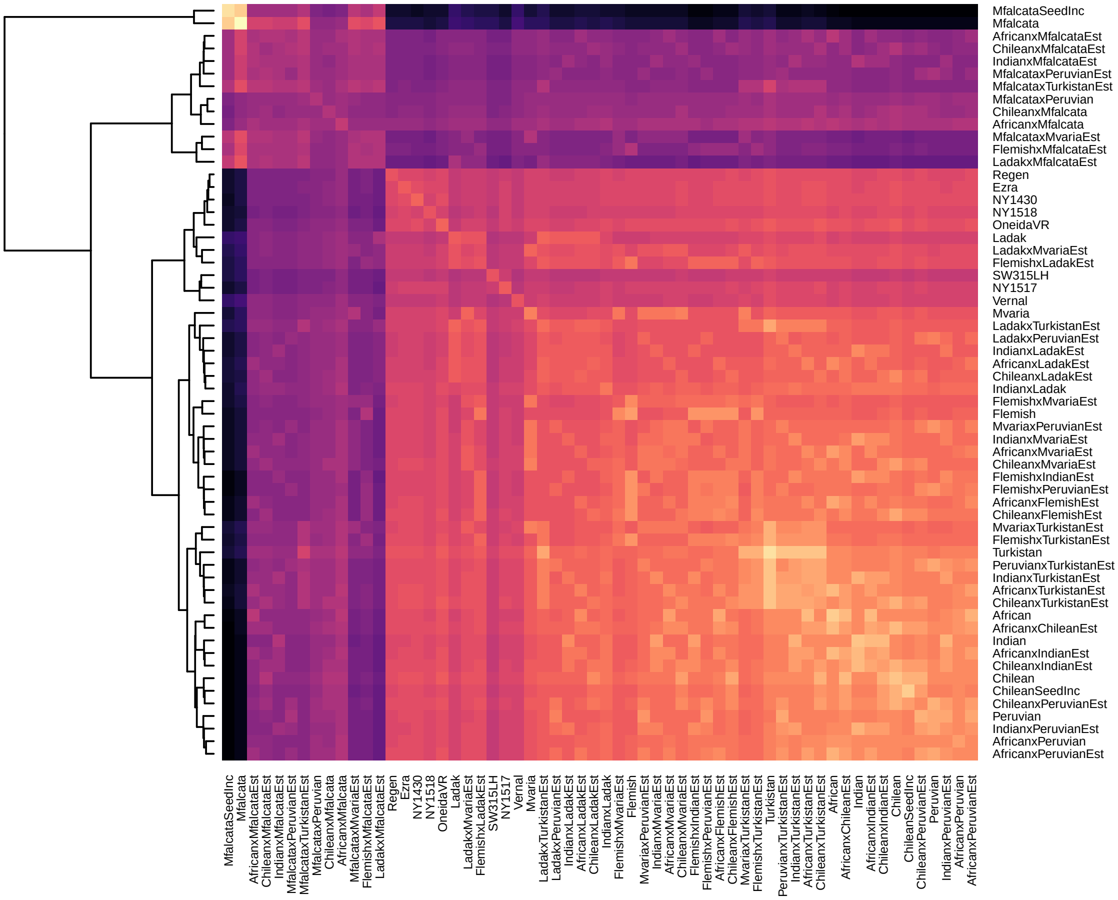


Figure 1: Heatmap of pairwise F_{st} values between populations, with rows and columns sorted using hierarchical clustering. Brighter colors indicate higher F_{st} values, and therefore higher additive genetic covariance between populations.

Initial results confirm hybrid parentage and the ability to construct hybrid genotypes *in silico*, as well as the confirmation of biological replicates as controls (Figure 1).

3.2 Efficacy of pop-level Genotyping approach

A total of 77,688,674 polymorphic sites were identified in the panel of alfalfa populations. Filtering to keep sites with at least 20, but no more than 125 reads for each population produced a total of 273,939 sites. These were filtered to obtain total allele frequencies of $0.05 < p < 0.95$, resulting in 89908 sites for estimating genetics relationships across populations.

Of the two methods used to estimate genetic relationships, the simple covariance of allele frequencies tended to be more stable for model fitting, suggesting that estimating the gentic covariance based on pairwise F_{st} statistics does not produce a positive definite matrix. This is likely due to the unknown (i.e. because the reference (or basal?) Fst cannot be estimated due to lack of degrees of freedom, the matrix has a rank ¡ n (likely n-1 or 2)) limitations of.

Further investigation is required to determine how this arises.

Correlations between expected and empirical allele frequency estimates for the diallel popultion hybrids was very high, ranging from 0.9 to 0.9, further validating the ability of sequenced based methods to accurately sample alleles and estimate frequency, even in bulk samples with unbalanced numbers of cells per individual, given sufficient individuals are included inthe bulk (i.e. large sample size).

Genetic relationships were similar to those estimated in the same population lusing AFLPs (Segovia Lerma 2004), with families clustering together (see figure X), especially

The high level of genetic variability in the diallel was demonstrated by high genomic predictability, with leave one family out genomic prediction accuracies ranging from 0.55-0.97 (Figure XXX). Generally, inclusion of marker information increased prediction accuracy compared to a pedigree based prediction model. Inclusion of dominance predictors also increased accuracy for several families, demonstrating that dominance effects are important contributors, and will allow for prediction of hybrid vigor. This data set is atypical in the genetic signal due to the large genetic variability and known recent familial relationships, but we chose this population as a proof of concept because it is small (low cost), and has a high potential to detect differences (high genetic variance).

3.3 Genetic covariance between populations

Pairwise F_{st}'s were used to calculate additive genetic relationships between populations (Weir and Hill 2002). These were then used to evaluate genomic predictive ability of yearly dry matter forage yield in the diallel population which was previously evaluated in Las Cruces, NM in 1997 and 1998 (Segovia-Lerma et al. 2004). F_{st}'s showed increased overall predictive ability over pedigree or dominant markers (AFLPs), especially for the most highly unrelated family. This suggests that tracking allele frequencies at many loci better captures relationships between sites and causal loci, rather than simply tracking familial relationships (Figure 2). We have yet to evaluate predictive ability for the populations currently under evaluation, but will be doing so in the near future.

Table 1: Leave one out genomic prediction accuracy for F_st and allele freq cov in Geneva trial.

Harvest	Year	Fst	covAF
2	2019	0.27	0.43
3	2019	0.75	0.73
1	2020	0.51	0.24
2	2020	0.94	0.97
sum	0.90	0.79	

[paragraph about comparison in Fst being better than simple covariance]

3.4 Phenotypic description

[paragraph noting the genetic variability of lack thereof in harvests for end use traits]

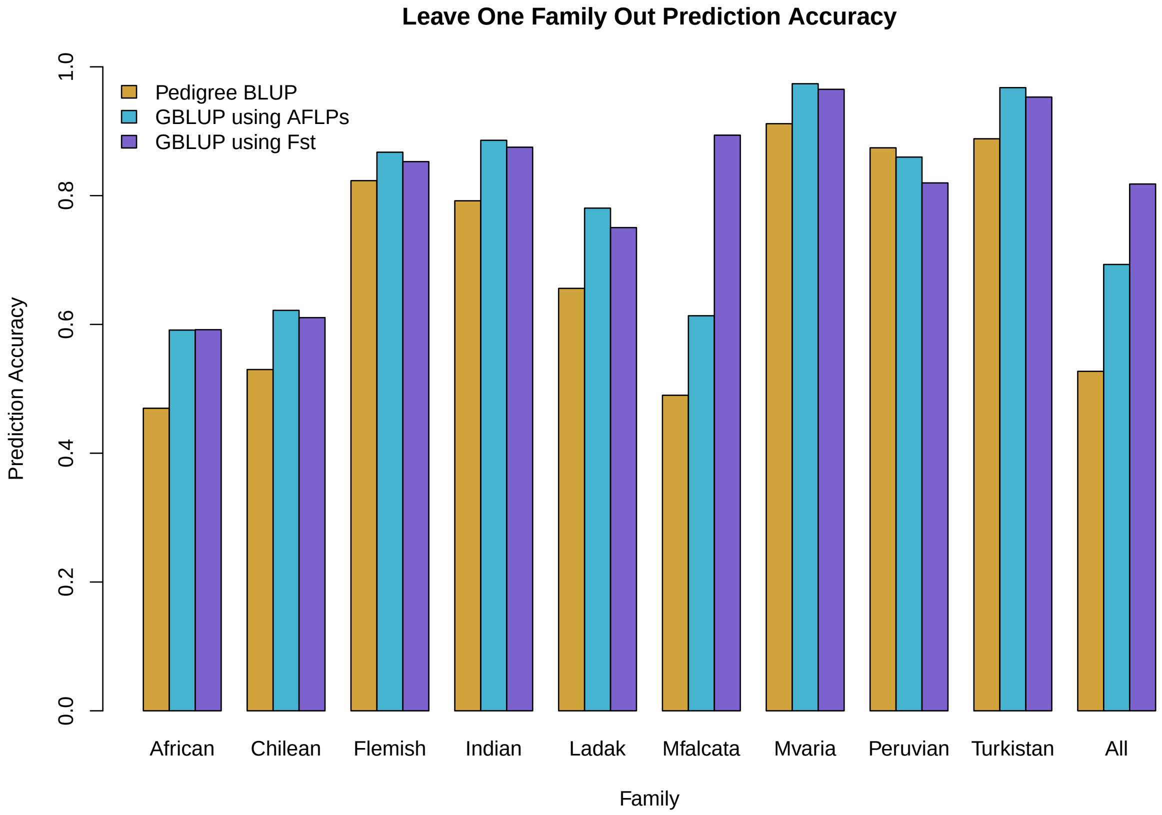


Figure 2: Prediction accuracy using a leave one family out strategy for a diallel population with 9 parental populations, and 36 hybrid populations of alfalfa. For each of the nine parents, all entries with that parent were removed and predicted using the remaining eight families and the additive genetic covariance estimated using pedigrees, dominant markers (AFLPs; Segovia-Lerma et al. 2004), or F_{st} statistics calculated from variant frequencies determined by whole-genome resequencing.

3.5 Genetic parameters

[Add quality tables here for harv 2 , 2019 and 2020]

3.6 Genetic Relationships between Vegetative Indices and forage yield and quality

- Yield * Genetic correlations very high

  - Quality * no genotype effects * estimates close to zero

  -

  [insert figure of genetic correlation to VIs here!]

  [Paragraph about correlations of VIs with yield, quality.]

3.7 Using Vegetative indices to predict unobserved phenotypes

- predicting harvests with vegetative indices

Table 2: Trait means, heritability and ANOVA p-values for forage yield, crude protein and neutral detergent fiber across 2 harvests in each of 2019 and 2020 for all fifteen entries, or only the eight entries with genotypic information.

trait	year	harvest	mean kg ha^{-1}	iidh2	iidh28	covh28	anovaPval	anovaPval8
Forage yield	2019	2	1.51	0.11	0.11	0.00	0.1073	0.1751
Crude Protein	2019	2	0.18	0.07	0.04	0.00	0.1888	0.3219
NDF	2019	2	0.47	0.15	0.14	0.1200	0.0478	0.1196
Forage yield	2019	3	0.99	0.21	0.31	0.40	0.0117	0.0110
Forage yield	2020	1	2.65	0.40	0.24	0.24	< 0.0001	0.0326
Forage yield	2020	2	1.44	0.71	0.63	0.71	< 0.0001	< 0.0001
Crude Protein	2020	2	0.24	0.42	0.43	0.57	< 0.0001	0.0012
NDF	2020	2	0.43	0.00	0.00	0.00	0.7521	0.7195

Table 3: Genetic correlations of forage yield by harvest. Above diagonal is genetic correlations, below diagonal is error correlations, and diagonal is heritability

	yld2_19	yld3_19	yld1_20	yld2_20
yld2_19	0.11	0.63	0.85	0.71
yld3_19	0.86	0.21	0.88	0.62
yld1_20	-0.26	-0.14	0.40	0.74
yld2_20	-0.06	0.05	0.35	0.71

- predicting values without harvesting all reps. * genetic relationships allow for reduced phenotyping * unclear if veg indices

[two paragraphs about how VIs do increase prediction, but most gain comes from including genetic covariance!]

3.8 Genotype specific Growth curves

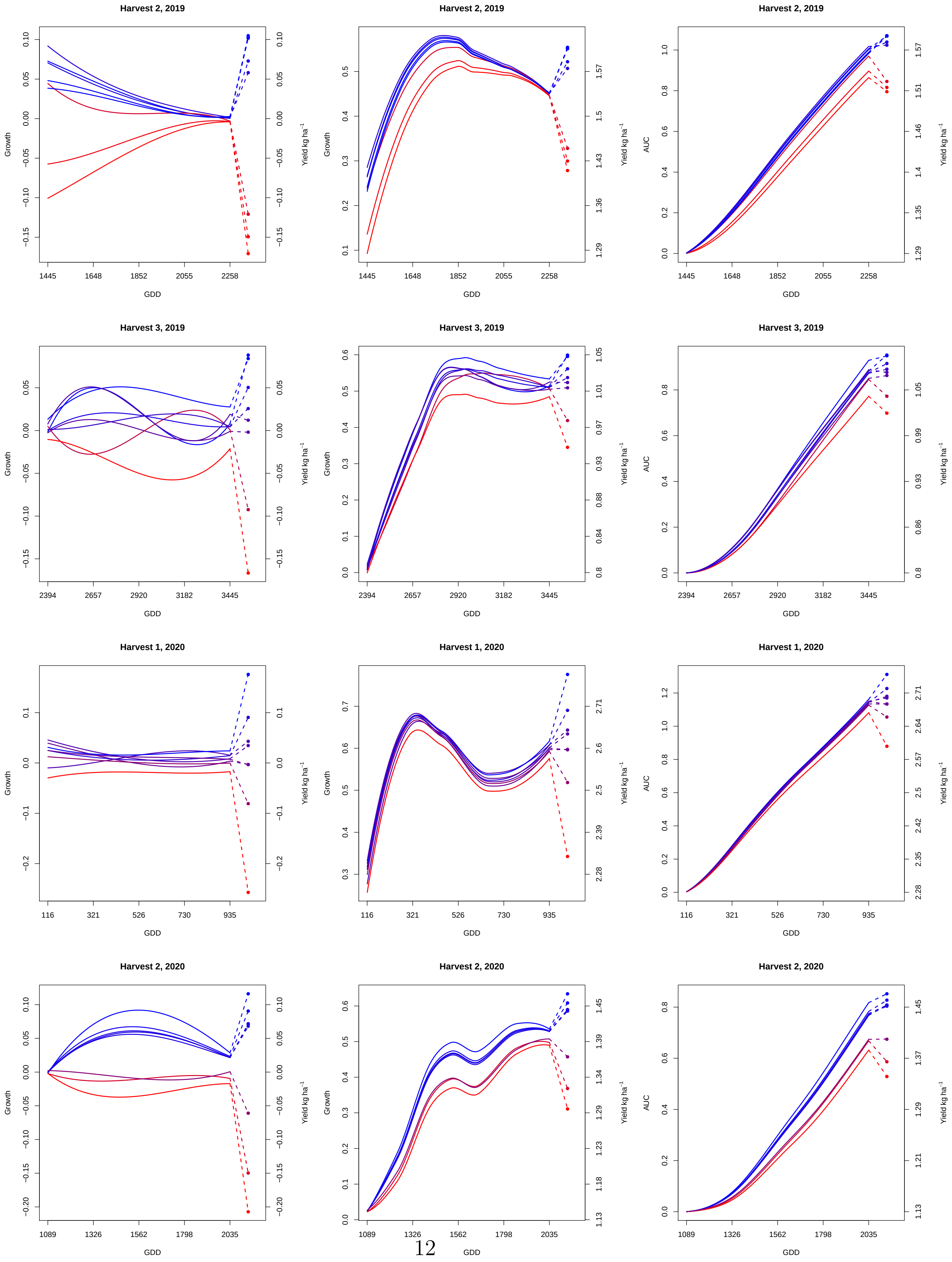
- Growth curves and relationships between curve and forage yield

- Area under the curve and relationship to forage yields, quality

4 Conclusion

- Other linear and non-linear indices should be investigated - larger sample sizes needed to validate and decipher more consistent trends - inclusion of genetic covariance important for prediction, perhaps more than VIs?

This research was used as preliminary evidence for submission of a grant proposal to FFAR: Seeding Solutions with a combined budget of $767,605, which will include a larger three year study of 24 alfalfa varieties evaluated at Cornell, and 24 varieties evaluated for two years at New Mexico State University.



Harvest 2, 2019
Growth
Yield kg ha^{-1}
GDD
1445
1648
1852
2055
2258
0.10
0.05
0.00
−0.05
−0.10
−0.15
Harvest 2, 2019
Growth
Yield kg ha^{-1}
GDD
0.1
0.2
0.3
0.4
0.5
1.29
1.36
1.43
1.5
1.57
Harvest 2, 2019
AUC
Yield kg ha^{-1}
GDD
0.0
0.2
0.4
0.6
0.8
1.0
1.29
1.35
1.4
1.46
1.51
1.57
Harvest 3, 2019
Growth
Yield kg ha^{-1}
GDD
2394
2657
2920
3182
3445
0.05
0.00
−0.05
−0.10
−0.15
Harvest 3, 2019
Growth
Yield kg ha^{-1}
GDD
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
0.84
0.88
0.93
0.97
1.01
1.05
Harvest 3, 2019
AUC
Yield kg ha^{-1}
GDD
0.0
0.2
0.4
0.6
0.8
0.8
0.86
0.93
0.99
1.05
Harvest 1, 2020
Growth
Yield kg ha^{-1}
GDD
116
321
526
730
935
0.1
0.0
−0.1
−0.2
Harvest 1, 2020
Growth
Yield kg ha^{-1}
GDD
0.3
0.4
0.5
0.6
0.7
2.28
2.39
2.5
2.6
2.71
Harvest 1, 2020
AUC
Yield kg ha^{-1}
GDD
0.0
0.2
0.4
0.6
0.8
1.0
1.2
2.28
2.35
2.42
2.5
2.57
2.64
2.71
Harvest 2, 2020
Growth
Yield kg ha^{-1}
GDD
1089
1326
1562
1798
2035
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Harvest 2, 2020
Growth
Yield kg ha^{-1}
GDD
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1.13
1.18
1.23
1.29
1.34
1.39
1.45
Harvest 2, 2020
AUC
Yield kg ha^{-1}
GDD
0.0
0.2
0.4
0.6
0.8
1.13
1.21
1.29
1.37
1.45

Table 4: Prediction accuracy of unobserved harvests

	VI	PC	K	iid
yld1_20_yld2_201	0.55	0.74	0.53	0.49
yld1_20_yld2_202	0.90	0.94	0.76	0.71
yld3_19_yld2_201	0.68	0.73	0.86	0.81
yld3_19_yld2_202	0.90	0.94	0.82	0.74
yld3_19_yld1_201	0.68	0.73	0.93	0.94
yld3_19_yld1_202	0.55	0.74	0.64	0.61
yld2_19_yld2_201	0.81	0.71	0.68	0.68
yld2_19_yld2_202	0.90	0.94	0.88	0.84
yld2_19_yld1_201	0.82	-0.71	0.72	0.72
yld2_19_yld1_202	0.55	-0.74	0.65	0.64
yld2_19_yld3_191	0.82	0.71	0.55	0.54
yld2_19_yld3_192	0.68	0.73	0.76	0.73
yld2_20	0.90	0.94	0.84	0.80
yld1_20	0.55	0.74	0.62	0.60
yld3_19	0.68	0.73	0.88	0.87
yld2_19	0.81	0.71	0.66	0.67

5 Acknowledgments

We want to express special appreciation to the Noble Research Institute for providing the alfalfa genome assembly, without which this work would not have been possible. In particular to we would like to recognize Michael Udvardi and Maria Monteros for providing the data, Andrew Segna for working to establish an MTA and to Andrew Farmer, who was integral to the graciously worked through several early alignment questions with us. We are grateful to Don Viands and Julie Hansen and the rest of the the Forage Breeding and genetic program team at Cornell university, for allowing us to image and genotype their materials. We hope that this collaboration is sustained well into the future, with the

References

Barnes, DK et al. (1977). "Alfalfa germplasm in the United States: genetic vulnerability, use, improvement, and maintenance." In: *USDAARS Technical Bulletin* 1571, pp. 1–21.

Li, Heng (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27.21, pp. 2987–2993.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14, pp. 1754–1760.

Misztal, I et al. (2002). "BLUPF90 and related programs (BGF90)". In: *Proceedings of the 7th world congress on genetics applied to livestock production.* Vol. 33, pp. 743–744.

Morales, Nicolas et al. (2020). "ImageBreed: Open-access plant breeding web–database for image-based phenotyping". In: *The Plant Phenome Journal* 3.1, e20004.

Segovia-Lerma, A et al. (2003). "AFLP-based assessment of genetic diversity among nine alfalfa germplasms using bulk DNA templates". In: *Genome* 46.1, pp. 51–58.

Segovia-Lerma, A et al. (2004). “Population-based diallel analyses among nine historically recognized alfalfa germplasms”. In: *Theoretical and applied genetics* 109.8, pp. 1568–1575.

Sharratt, BS, CC Sheaffer, and DG Baker (1989). “Base temperature for the application of the growing-degree-day model to field-grown alfalfa”. In: *Field Crops Research* 21.2, pp. 95–102.

Weir, Bruce S and William G Hill (2002). “Estimating F-statistics”. In: *Annual review of genetics* 36.1, pp. 721–750.

Table 5: Accuracy of genetic values when 1 to 4 replications of forage biomass data are set to missing.

	pred	hidden reps	withNDVI	NDVIfixed	noNDVI	iid
yld2_19	PC	1	0.66	0.61	0.92	0.94
yld3_19	PC	1	0.96	0.95	0.97	0.97
yld1_20	PC	1	0.89	0.94	0.95	0.96
yld2_20	PC	1	0.99	0.86	0.99	0.99
yld2_19.1	PC	2	0.51	0.51	0.87	0.85
yld3_19.1	PC	2	0.92	0.89	0.93	0.92
yld1_20.1	PC	2	0.86	0.89	0.90	0.90
yld2_20.1	PC	2	0.98	0.81	0.97	0.97
yld2_19.2	PC	3	0.58	0.41	0.75	0.72
yld3_19.2	PC	3	0.86	0.81	0.88	0.84
yld1_20.2	PC	3	0.84	0.79	0.82	0.80
yld2_20.2	PC	3	0.97	0.74	0.95	0.94
yld2_19.3	PC	4	0.28	0.24	0.53	0.49
yld3_19.3	PC	4	0.79	0.62	0.79	0.76
yld1_20.3	PC	4	0.75	0.68	0.73	0.66
yld2_20.3	PC	4	0.95	0.64	0.80	0.80
yld2_191	VI	1	0.83	0.34	0.92	0.94
yld3_191	VI	1	0.90	0.85	0.97	0.97
yld1_201	VI	1	0.84	0.92	0.95	0.96
yld2_201	VI	1	0.99	0.77	0.99	0.99
yld2_19.11	VI	2	0.72	0.34	0.86	0.84
yld3_19.11	VI	2	0.86	0.80	0.93	0.92
yld1_20.11	VI	2	0.71	0.83	0.90	0.90
yld2_20.11	VI	2	0.98	0.73	0.97	0.97
yld2_19.21	VI	3	0.67	0.28	0.75	0.71
yld3_19.21	VI	3	0.82	0.73	0.88	0.84
yld1_20.21	VI	3	0.61	0.66	0.82	0.80
yld2_20.21	VI	3	0.96	0.69	0.95	0.94
yld2_19.31	VI	4	0.60	0.26	0.53	0.49
yld3_19.31	VI	4	0.70	0.61	0.72	0.58
yld1_20.31	VI	4	0.26	0.43	0.73	0.66
yld2_20.31	VI	4	0.92	0.52	0.86	0.86