From Research Findings to Educational Frameworks: Breeding Progress and Genomic Insights into Intermediate Wheatgrass at the University of Minnesota

A Dissertation

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Hannah Kay Stoll

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

James A. Anderson

2024

Dedication

This work is dedicated to the friends, family, and loved ones who have encouraged, prayed for, and supported me throughout this journey… even though they will likely never read this dissertation!

Abstract

Intermediate wheatgrass (IWG, *Thinopyrum intermedium* (Host) Barkworth & D.R. Dewey) is a perennial grain crop that has been under development for inclusion in continuous living cover systems worldwide since the 1980s and specifically at the University of Minnesota (UMN) since 2011. IWG offers multiple ecosystem benefits as a perennial grain crop and has the potential to support rural economies by offering farmers a high-value alternative to the relatively few annual crop options that dominate markets and the landscape. However, improving grain traits, developing markets, and establishing optimal agronomic practices are essential for establishing and ensuring the long-term sustainability of IWG as a grain crop in the Midwest and beyond. This dissertation sheds light on the amount of breeding progress and underlying genomic regions for key IWG traits at UMN and uses this knowledge to make recommendations for future breeding efforts. Thus, 242 parent genets (genetically unique individuals) from UMN IWG breeding cycles 2, 3, 4, and 5 were evaluated in 2 locations (St. Paul and Lamberton, MN) in 2021 and 2022. Genets were genotyped and phenotyped for a variety of domestication and agronomic traits including shattering, brittle rachis, seed size, seed weight, spike characteristics, anthesis timing, and plant height. This dissertation demonstrates that the rate of genetic gain for several traits is significant and improving across breeding cycles; floret and average shattering decreased by ~5% each cycle, and seed area increased by 1% each cycle. Moreover, 33 quantitative trait loci (QTL) for these traits were identified in a Genome-Wide Association Study (GWAS), individually explaining 13% of phenotypic variation, on average. Some QTL appeared to be in close proximity to previously identified domestication – related genes from other studies, and favorable allele frequencies for some QTL appeared to increase across cycles. With these findings, we conclude that the rate of IWG breeding progress could be increased by improved

phenotyping methods, incorporation of identified QTL in genomic selection procedures, and decreased time per breeding cycle. In addition to exploring IWG breeding progress at UMN, this dissertation explores the potential to disseminate scientific research into relevant science curriculum. Here, concepts such as the integration of quantitative skills and active learning into curriculum are explored, and a framework for developing and evaluating these curricula is proposed.

Table of Contents

1. Assessing Genetic Gain in an Intermediate Wheatgrass Improvement Program: A Retrospective Analysis

1.1 Abstract

The perennial crop intermediate wheatgrass (IWG) has been under development as a grain crop since 2011 at the University of Minnesota (UMN). Breeding efforts at UMN have targeted larger seed size, reduced shattering, and improved threshability, among other traits. Calculating the rate of genetic gain provides insights into breeding progress and informs adjustments in future breeding practices. In this study, materials from multiple breeding cycles of the UMN IWG breeding program were grown at the same locations for multiple years to estimate realized genetic gain of domestication and agronomic traits. A set of 242 parents from breeding cycles 2, 3, 4, and 5 were planted in an augmented experimental design and evaluated over 2 years in 2 locations. Improvement in the mean values for selected traits from cycle 2 to cycle 5 was observed for most traits. Regression analysis used to estimate the rate of genetic gain was significant for both floret and average shattering, spikelet density, seed area and width, and 10-spike seed weight and spike weight. Floret and average shattering decreased by ~5% each cycle, and seed area increased by 1% each cycle. With this information paired with calculated heritability values, we recommend phenotyping the two types of shattering independently and selecting earlier-anthesis genotypes to improve yield component traits. We conclude that progress for several traits may take decades to achieve desired population means, and could be improved with better phenotyping methods, genomic selection applications, and decreased time per breeding cycle. These findings are applicable to other IWG breeding programs and perennial grain domestication programs.

1.2 Introduction

Intermediate wheatgrass (IWG, *Thinopyrum intermedium* (Host) Barkworth & D.R. Dewey) is a perennial grain crop under development for inclusion in continuous living cover systems across the world. The need for productive and sustainable agricultural systems drives the development of perennial grain crops, IWG included. Perennial crops significantly reduce nutrient leaching and mitigate subsequent water quality issues (Pimentel et al., 2012; Jungers et al., 2019). Moreover, the continuous living cover provided by perennial crops are proven to conserve soil and water resources, improve soil health, and reduce greenhouse gas emissions (Pimentel et al., 2012; Kaye and Quemada, 2017; Reilly et al., 2023).

IWG is a forage and grain crop with numerous uses. IWG is an allohexaploid (2n = 6x = 42), obligate outcrossing, highly heterozygous species with a large 12.75 Gb genome (Vogel et al., 1999). Grain of IWG can be marketed under the Kernza® trademark, with the owners of the trademark determining the criteria necessary for IWG grain to be marketed as such to ensure both grain quality and sustainability specifications (DeHaan and Ismail, 2017). Prior to development as a perennial grain crop for human consumption, IWG was used as a forage crop beginning in the early 20th century after its introduction to North America from Russia (Vogel and Jensen, 2001). Later, it was leveraged as a genomic resource for wheat (*Triticum aestivum* L.) improvement (Sharma et al., 1995; Fedak and Han, 2005). In the late 20th Century, The Rodale Institute (Kutztown, PA, USA) identified IWG as a promising candidate to be a perennial grain crop (Wagoner, 1990). The Rodale Institute completed 2 cycles of selection for spikelet fertility and seed size beginning in 1988 (DeHaan et al., 2014). Later, in 2001, The Land Institute (TLI, Salina, KS) used these materials to begin a breeding program (DeHaan et al., 2018). In 2011, The University of Minnesota (UMN, St. Paul, MN) acquired 2560 genets (or “genetically

unique plants", an ecology term now commonly used for IWG - Zhang et al., 2016) from TLI's phenotypic recurrent selection breeding program. Those 2560 genets were derived from 66 TLI Cycle 3 genets, which were crossed to form the 2560 individual plants, marking the inception of the first breeding cycle (UMN-C1) of the IWG breeding program at UMN (Zhang et al., 2016). In 2019, the first commercial variety, MN-Clearwater, a synthetic variety from UMN-C1, was released in the United States (Bajgain et al., 2020b). As of 2023, the UMN breeding program has completed 7 Cycles of recurrent selection for both agronomic and domestication traits.

Phenotypic recurrent selection originally informed selections at both The Rodale Institute and TLI, but both TLI and UMN now leverage genomic selection for population improvement (Zhang et al., 2016; Bajgain et al., 2020a; Crain et al., 2021b). Tools such as high throughput phenotyping and affordable genetic sequencing have and will continue to support rapid breeding improvements in IWG. Genotyping-by-sequencing, or GBS (Poland et al., 2012), is routinely used in the UMN's IWG breeding program for genomic prediction (Bajgain et al., 2019, 2020a; Crain et al., 2020; Bajgain et al., 2022b). Several agronomic and domestication traits are under selection for improvement. Since 2011, breeding targets have included agronomic traits such as heading date, plant height, grain yield, seed weight and size, spike weight, and plant biomass (Zhang et al., 2016; Bajgain et al., 2019). Additionally, domestication traits such as threshability (Zhang et al., 2016) and resistance to shattering were selected upon beginning with Cycle 1 and Cycle 2, respectively. Overall, the strongest selection intensity has been for larger seed size, reduced shattering, and improved threshability. Secondarily, plants with high grain yield, consistent heading and anthesis, i.e. most of the tillers on the spaced plants reach anthesis at the same time, and reduced lodging were selected (Bajgain et al., 2022a).

Reports of expected genetic gain in the context of improvements due to genomic selection for agronomic and domestication traits in IWG have been promising. For example, at UMN, varying genomic selection models have demonstrated the potential to maximize expected genetic gain for numerous traits (Bajgain et al., 2020a). Similarly, Crain et al., 2021b quantified the benefits of genome-wide prediction over phenotypic selection, for example, resulting in a 2.6 fold increase in expected genetic gain for spike yield using genomic selection.

Despite demonstrated progress from selection for agronomic and domestication traits (Bajgain et al., 2022a), a true estimate of realized genetic gain, or the observed gain due to selection over cycles has not been reported in IWG. Genetic gain has been widely reported for crops such as wheat (Sharma, et al., 2012) and corn (Duvick, 2005), and quantification of genetic gain in the context of genomic selection has been estimated in both wheat (Tessema et al., 2020) and corn (Zhang et al., 2017b) as well. The lack of genetic gain reporting in IWG is largely due to both how recently breeding efforts began in IWG and the difficulty of estimating realized genetic gain in a breeding population of a perennial outcrossing species. The time required to develop and release a variety from a breeding program spans years, usually decades, making it difficult to accurately compare individuals across breeding cycles. For example, comparing breeding program data collected in 1970 with genotype means from the same breeding program in 2020 would be difficult, as non-genetic factors such as management or climate differences inevitably confound these comparisons. One demonstrated way to estimate the rate of genetic gain of a breeding program across many years is to evaluate several released varieties from many years in the same environment in what's called an "Era Trial," as seen in Donmez et al (2001). While informative, this method likely doesn't provide an entirely accurate estimate of genetic gain, since released varieties are not necessarily representative of all individuals in the breeding

program. In the case of IWG, there is only one commercially available released variety in the United States, MN-Clearwater, so an Era Trial would not be appropriate in any case. The best way to combat these confounding effects is to grow a sample of individuals from several breeding cycles representing the breadth of diversity in each cycle in the same set of environments. Then, the rate of realized genetic gain can be estimated as the slope of the regression line of the mean breeding value for each cycle (Eberhart, 1964; Rutkoski, 2019).

The objectives of this study were to 1) estimate realized genetic gain for key domestication and agronomic traits in UMN's IWG domestication program; and 2) utilize this knowledge to make recommendations for future breeding practices to maximize genetic gain for these traits.

1.3 Materials and Methods

1.3.1 UMN IWG Breeding Population

UMN's IWG breeding population originated from 66 Cycle 3 genets from TLI. Figure 1.1 depicts the breeding program in visual format. The first breeding cycle (hereby UMN-C1) contained 2560 genets resulting from crosses of 66 TLI-C3 genets. UMN-C1 was planted in September 2011 in St. Paul, MN, and plant traits were recorded in both 2012 and 2013. UMN-C1 plants were evaluated for heading date, plant height, spike weight, grain yield, threshability, seed weight, and plant biomass. From UMN-C1, 50 of the best plants were selected as UMN-C2 parents by genomic selection-based phenotypic selection for the aforementioned traits (Zhang et al., 2016). UMN-C2 parents were crossed, resulting in 1,660 UMN-C2 progeny, which were transplanted into a St. Paul, MN field in fall 2013 and evaluated in 2014 and 2015. Seventy-two UMN-C2 genets were selected (now UMN -C3 parents) from genomic selection models trained

with the 2014-2015 UMN-C2 phenotype data. UMN-C3 parents were crossed, resulting in 608 individual plants (8 from each mother plant except for four mother plants which had 16 plants each). UMN-C3 was planted in 2016 and evaluated in 2017 and 2018 at two MN locations: St. Paul and Crookston (Bajgain et al., 2019). UMN-C4 was initiated by crossing 73 UMN-C3 genets that were predicted by a genomic selection model trained on UMN-C3 data to be the best performers for grain yield, seed size, shatter resistance, free threshing, and plant height. A population of 657 genets (9 from each parent) was established in September 2018 in St. Paul and Crookston and evaluated during 2019-2020. Likewise, UMN-C5 was initiated by crossing 40 UMN-C4 genets that were predicted by a genomic selection model trained on UMN-C4 data to be the best performers for grain yield, seed size, shatter resistance, free threshing, and plant height. A population of 684 genets (~17 from each parent) was established in September 2020 in St. Paul and Crookston and evaluated during 2021-2022.

1.3.2 Plant Materials and Experimental Design

Since IWG is an obligate outcrossing species, vegetative cloning is necessary for replication of genets. Thus, to perpetuate true replicates of each parent genet, parent plants from UMN-C2 through UMN-C4 were cloned and maintained in a living nursery in a field in St. Paul, MN since the outset of the IWG breeding program at UMN. The 66 UMN-C1/TLI-C3 parents were not available from TLI (only TLI-C3 progeny were shared), thus, UMN-C1 parents were not included in this study. For this experiment, selected parents of UMN-C2, C3, and C4 were dug from the living nursery and cloned in 2020. Additionally, clones of UMN-C5 parent plants were obtained by cloning plants that had been recently selected and grown in the greenhouse in 2020. This resulted in 242 genets from 4 cycles, which were cloned to be planted at 2 locations: St. Paul, MN and Lamberton, MN in August 2020. Plants were phenotyped in 2021 and 2022.

Parent plants were planted in an augmented randomized complete block design (ARCBD) with 8 incomplete blocks (iblock) with 33 entries plus 3 unique checks per iblock for a total of 36 genets per iblock and 266 total entries. Check genets were randomly chosen and vegetatively cloned from the parent nursery (but were not parents themselves). The ARCBD is a unique case of the partially balanced incomplete block design (Cochran and Cox, 1957).

1.3.3 Phenotyping Methods

Plants were phenotyped in 2021 and 2022 for key domestication and agronomic traits. Those included height, anthesis date, brittle rachis shattering, floret shattering, average shattering, threshability, spikelet density, 10-spike dried weight, 10-spike seed weight, thousand grain weight (TGW), seed area, seed width, seed length, and length:width ratio. Height was measured from the base of the plant to the tallest spike in a space plant in centimeters. Anthesis was recorded in Julians (days after January 1st) when approximately 50% of a plant was actively shedding pollen.

For postharvest traits, 10 spikes deemed representative of the whole plant were harvested from each plant when the majority of the experiment was at optimal harvest timing and carefully stored and dried to minimize shattering prior to phenotyping. Ten spikes were weighed to get 10-spike weight, then, six random spikes were measured for spike length and spikelet number; spikelet number was divided by spike length to estimate spikelet density. Shattering was separated into two types: brittle rachis shattering and floret shattering, since the two have previously been proven to be independently genetically controlled (Altendorf et al., 2021b). Additionally, an “average shattering” rating is reported, which is simply the average of floret and brittle rachis shattering, as this type of shattering rating was used for selections in the IWG

breeding program. For shattering, spikes were dropped from a consistent one-meter height onto a metal lab bench and then rated for brittle rachis and floret shattering on a 0 to 9 scale. For floret shattering, a rating of 0 being 0% shattering, 1 being 1%-10% shattering, and 9 being 91-100% shattering. For brittle rachis shattering, 0 being no rachis breaks, 1 being 1 rachis break, and 9 being 9 or more rachis breaks.

The 10-spike samples were threshed using a Wintersteiger LD 350 (Wintersteiger Inc, Salt Lake City, USA) and weighed. Samples were then rated for threshability (or a sample hulled vs. naked seed proportion) using a visual 0-9 rating scale with 0 being 0-10% of the sample were naked seeds, 9 being 91-100% of seeds were naked. From here, 10-spike seed weight in grams was recorded, and a subsample of 50-150 naked seeds was taken for thousand grain weight (TGW), seed area, seed width, seed length, and seed length:width ratio on the Marvin Optical Seed Analyzer (GTA Sensorik GmbH, Germany).

1.3.4 Statistical Analysis

Adjusting each incomplete block based on check genet values, per the planned experimental design (augmented randomized complete block design) did not accurately capture the heterogeneity of experimental units due to spatial variation in the field. Subsequently, a model using a residual structure variance-covariance matrix (AR1 x AR1) to model spatial dependency in combination with and without a genomic relationship matrix (GRM) was fit to adjust phenotypic values in concordance with spatial variation and relatedness (a technique successfully applied by Crain et al. (2021b). However, it was found that a simpler mixed model without accounting for spatial dependency or genomic relatedness was the better model fit by AIC value and likelihood ratio tests and did not result in over-shrinkage of genotypic effects.

Best linear unbiased estimates (BLUEs) for each breeding cycle were calculated using separate mixed models for each trait. Cycle was considered a fixed effect, and genet nested within cycle and environment were fit as random effects, with trait as the response variable. The model was

$$Y_{ijk} = \mu + Cycle_{ij} + Genet_{ijk} + Environment_{jk} + \varepsilon_{ijk}$$

Where Y_{ijk} is the observed value for trait at the *i*-th level of Cycle, *j*-th level of Genet within the *i*-th level of Cycle, and *k*-th level of Environment, μ is the intercept, $Cycle_{ij}$ is the fixed effect of the *i*-th level of Cycle, $Genet_{ijk}$ is the random effect of the *j*-th level of Genet within the *i*-th level of Cycle, $Environment_{ik}$ is the random effect of the *k*-th level of each environment, and ϵ_{ijk} is the random error $N(0, \sigma^2)$. Mixed model analyses were performed using the lme4 package with R software (Bates et al., 2015; R Core Team, 2021).

After the application of the mixed model, adjusted Cycle means and regression coefficients for the fixed effect of Cycle were obtained. The slope can be interpreted as the amount of gain per cycle for a specific trait (Rutkoski, 2019), and the intercept can be interpreted as the estimated mean at Cycle 0. The relative gain per cycle is estimated as the ratio between the slope and intercept as a percent (Table 1.1). In addition, Pearson correlation coefficients for all traits were calculated using genet means across all cycles and for each cycle independently (Figure 1.3).

Genet-mean heritability in the broad sense (H^2) was estimated as:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{Env}^2}{E} + \sigma_{Error}^2}$$

Where $\boldsymbol{\sigma}_G^2$ is the genetic variance, $\boldsymbol{\sigma}_{Env}^2$ is the environmental variance, $\boldsymbol{\sigma}_{Error}^2$ is the residual error variance, and E is the number of individual environments (each year by location combination is considered a different environment).

1.4 Results

Simultaneously selecting for numerous agronomic and domestication traits each cycle theoretically should lead to improvements for the traits in which selection is being performed. In the final analysis, 225 genets were used due to some of the original 242 dying post-transplanting. In general, the trend for improvement of most traits was in the desired direction (Table 1.1, Figure 1.2). Moreover, it was verified that the trend for realized gain for each cycle was significant for floret and average shattering, spikelet density, seed area and width, and both 10-spike seed weight and spike weight (Table 1.1).

Broad sense trait heritability estimates (H^2) were moderate for all shattering traits, with values of 0.45, 0.58, and 0.42 for brittle rachis, floret, and average shattering, respectively (Table 1.1). For both floret shattering and average shattering (which is an average of floret and brittle rachis shattering), shattering decreased with each cycle. The relative gain for each cycle for floret shattering was -5.66% ($p < 0.050$) and -4.27% for average shattering ($p < 0.033$, Table 1.1). Moreover, threshability, another important domestication trait, was rated on a 0 to 9 scale with a higher number indicating more desirable free-threshing seed in the sample. Broad sense heritability for threshability was moderately high ($H^2 = 0.57$, Table 1.1). Although cycle was not

statistically significant in the mixed model regression ($p < 0.130$), there was an average increase of 3.64% per cycle (Table 1.1, Figure 1.2), and a notable sharp increase in the mean value from 4.66 in Cycle 4 to 5.53 in Cycle 5 (Supplemental Table S1).

We noted consistent improvements in cycle means and gain per cycle for traits including thousand grain weight (TGW), area, width, and length over successive cycles. Conversely, there was a favorable decrease in the length:width ratio observed across the cycles. Broad sense heritabilities for seed size traits were moderate to low, with the lowest heritability for seed width ($H^2 = 0.08$), and the highest heritability for TGW ($H^2 = 0.43$, Table 1.1). We found that the relative gain for each cycle for seed area was 1.17% ($p < 0.099$) and 0.79% for seed width ($p < 0.014$, Table 1.1). These results indicate that selection for larger seeds has been successful, albeit slow. Unexpectedly, 10-spike seed weight and spike weight trends also decreased by 3.32% and 1.93%, respectively (Table 1.1).

Pearson correlation coefficients for all traits provide insights into potential dependencies among traits. Interestingly, floret shattering was slightly positively correlated with TGW, seed area, width, and length ($r = 0.23$ to 0.32, Figure 1.3). Likewise, brittle rachis shattering, floret shattering, and average shattering were positively correlated with 10-spike seed weight ($r = 0.27$ to 0.42, Figure 1.3). This indicates that genets with higher (worse) values for shattering also tend to have larger seeds or higher 10-spike grain weight. However, when looking at trait correlations for each cycle independently, the correlation between any sort of shattering with seed size traits is only significant for cycles 2 and 3 ($p < 0.05$, Supplemental Figure S1a and S1b) and not significant for later cycles ($p < 0.05$, Supplemental Figure S1c and S1d). Moreover, across all cycles, floret shattering and average shattering were both moderately negatively correlated with anthesis date ($r = -0.43, -0.35$, Figure 1.3), suggesting later maturing genets shatter less. Lastly,

spikelet density was negatively correlated with TGW, seed area, and seed length (r = -0.23 to -0.27, Figure 1.3), suggesting that as spikelet density increases, TGW, seed area, and seed length decrease.

1.5 Tables and Figures

Figure 1.1: University of Minnesota Intermediate Wheatgrass Breeding Program. Cycles of selection, parents selected, resulting population sizes, and phenotypes evaluated are displayed. Planting and evaluation years refer to the progeny population, not the selected parents.

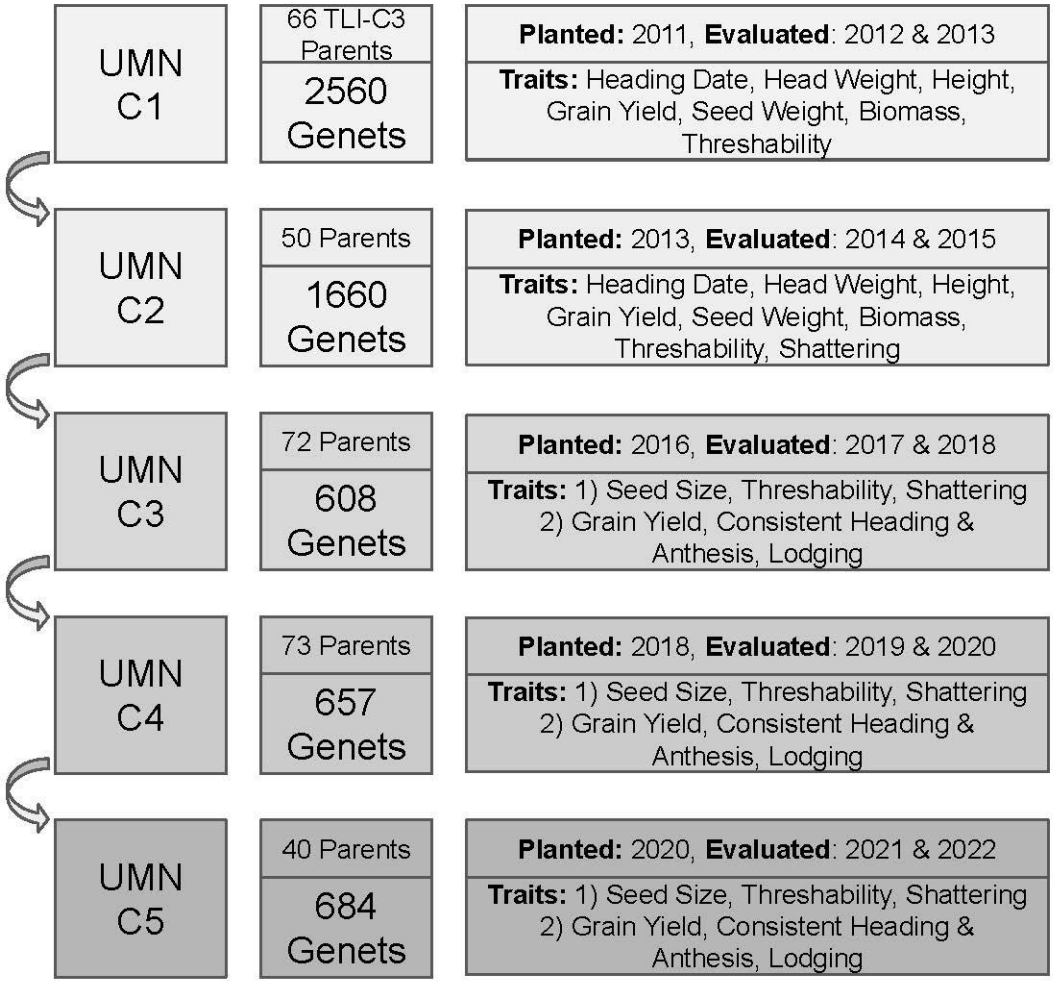


Figure 1.2: Regression line depicts progress for several traits across breeding cycles 2 through 5 of the University of Minnesota's intermediate wheatgrass breeding population. *, ** Cycle was significant in the regression analysis at the $p < 0.1$ and 0.05 levels, respectively.

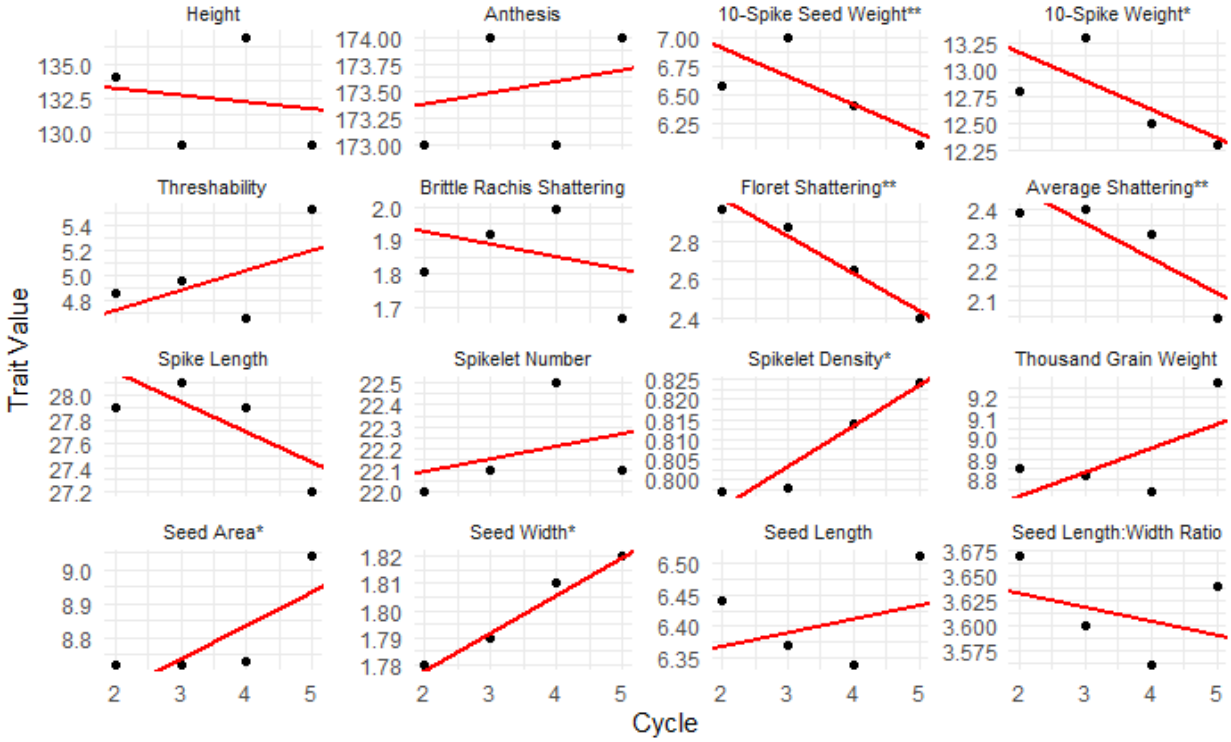
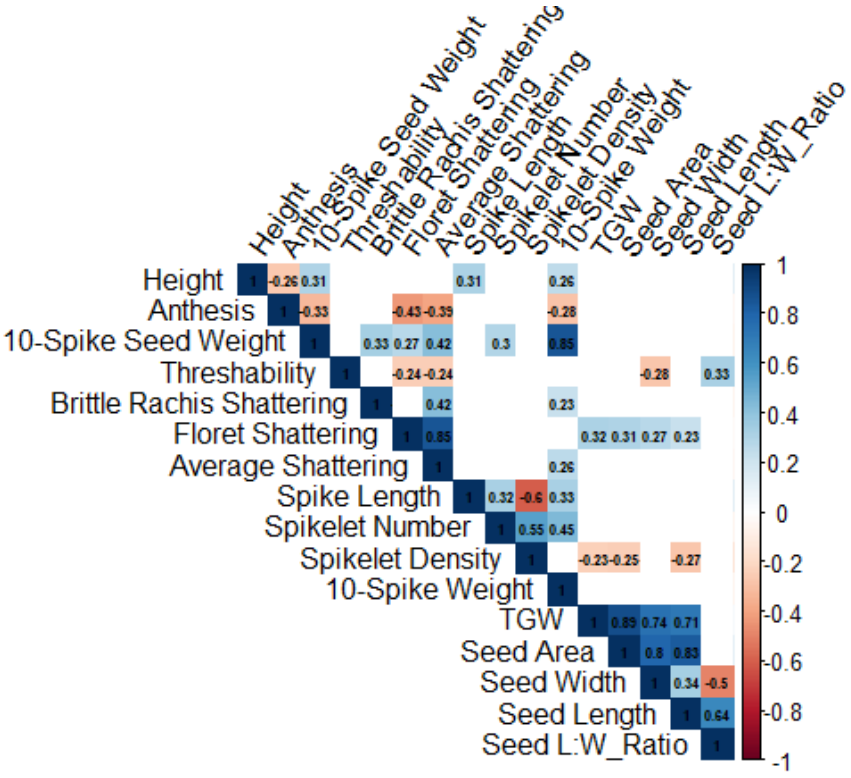


Table 1.1: Heritability and Parameter estimates for the linear regression of each trait for Cycles 2, 3, 4, and 5 of UMN breeding population. Gain per cycle (slope) estimates were significant at the α = .1* and .05** levels, respectively.

Trait	Intercept	T-Value	P-Value	Slope (Gain per Cycle)	Relative Gain	Broad-Sense Heritability
Height (cm)	134.23	-0.73	0.468	-0.507 NS	-0.38%	0.62
Anthesis (Julians)	173.18	1.07	0.284	0.105 NS	0.06%	0.18
10-Spike Seed Weight (g)	7.41	-2.24	0.026	-0.246**	-3.32%	0.39
Threshability	4.41	1.52	0.130	0.161 NS	3.64%	0.57
Brittle Rachis Shattering	2	-0.65	0.518	-0.038 NS	-1.88%	0.45
Floret Shattering	3.4	-1.97	0.050	-0.193**	-5.66%	0.58

Average Shattering	2.7	-2.15	0.033	-0.115**	-4.27%	0.42
Spike Length (cm)	28.69	-1.59	0.114	-0.251 NS	-0.87%	0.44
Spikelet Number	21.98	0.48	0.630	0.058 NS	0.26%	0.52
Spikelet Density	0.77	1.91	0.057	0.010*	1.30%	0.07
10-Spike Weight (g)	13.66	-1.90	0.059	-0.264*	-1.93%	0.33
TGW (g)	8.49	1.45	0.148	0.117 NS	1.38%	0.43
Seed Area (mm)	8.44	1.82	0.070	0.099*	1.17%	0.40
Seed Width (mm)	1.75	2.22	0.027	0.014**	0.79%	0.08
Seed Length (mm)	6.33	0.840	0.400	0.021 NS	0.34%	0.32
Seed L:W_Ratio (mm)	3.66	-0.91	0.363	-0.0142 NS	-0.39%	0.23

Figure 1.3: Pearson R Correlation Coefficients for traits significant at the $p < 0.05$ level, across Cycles 2, 3, 4, and 5 combined.



1.6 Discussion

This study evaluated the early stages of genetic improvement in a recurrent selection IWG breeding population over 5 cycles of genomic-assisted recurrent selection, which equates to nearly a decade of breeding work. Genomic selection's impact on genetic gain is well documented in animals and plants. The incorporation of genomic selection in breeding programs improves genetic gain by up to 160% in bovine dairy herds (Scott et al., 2021), up to 44.6% in radiata pine trees (McLean et al., 2023), and 14 – 50% for yield and stover-related traits in maize (Massman et al., 2013). Likewise, previous studies in IWG have demonstrated the ability of genomic selection models to increase expected genetic gain (Bajgain et al., 2020a; Crain et al., 2020, 2021b). However, no previous studies have calculated realized genetic gain in an IWG breeding program using plant populations grown in the same environments. Realized genetic gains for several traits can be considered satisfactory and indicative of UMN's success in selecting for several domestication and agronomic traits over just a few cycles of genomic-assisted recurrent selection. Moreover, the breeding program's primary emphasis on selecting for larger seed size, reduced shattering, and improved threshability is evident in this study. Notably, the genetic gain values calculated in this study are taken from a relatively small snapshot (4 cycles) in the grand picture of the numerous cycles and years necessary to achieve fixation for desirable domestication traits and improved agronomic traits.

In an ideal domesticated population, shattering should be non-existent (zero value). Both floret and average shattering significantly decreased each cycle. The per-cycle decrease for brittle rachis shattering was smaller and statistically insignificant, an expected trend since selection and phenotyping in the breeding program was only performed on an "average shattering" basis, rather than separating out the two types of shattering. If this rate of reduction in

shattering is continued at this rate, we estimate that floret shattering should be near zero by cycle 18, and average shattering score should be near zero by cycle 23. Recently, UMN switched to complete 1 cycle of selection each year, rather than 1 cycle every 2 years, meaning these shattering scores would be reached by 2034 and 2039, respectively. To date, selection for shattering resistance in the UMN breeding program is done using a general "average shattering" score that combines brittle rachis and floret shattering together into one score. In future breeding cycles, we recommend directly selecting on brittle rachis shattering in addition to floret shattering, which should reduce the amount of time to reach a zero average shattering score. These findings are in alignment with other studies, which suggest brittle rachis and floret shattering are controlled by different regions of the genome, thus should undergo independent phenotyping and selection (Altendorf et al., 2021b; Crain et al., 2022).

The correlation analysis conducted in this study reveals intriguing insights into the interplay between anthesis, shattering propensity, seed size, and 10-spike seed and spike weight traits in the breeding program. Anthesis date was negatively correlated with shattering scores, suggesting that later-maturing genets shatter less. In addition, brittle rachis, floret, and average shattering scores were all positively correlated with 10-spike seed weight, and floret shattering was positively correlated with seed size traits and TGW. This means that plants with a higher propensity for shattering also have larger seed size and 10-spike seed weights, all key yield component traits. This observation is expected, as a spike with higher seed mass would exert greater mechanical forces, thereby increasing the likelihood of shattering. Combining these findings with the anthesis trends, it is likely that plants that flower later may have less time to fill grain prior to harvest, leading to a lighter spike with smaller seeds and less shattering. Another likely possibility is that genets that flowered earlier are more physiologically mature at harvest,

and perhaps more likely to exhibit worse shattering. These results highlight the difficulty to improve multiple traits concurrently in a breeding program, especially when several of these traits are correlated with one another. Harvesting plants at staggered intervals optimized to anthesis date, i.e. as they reach optimal maturity, could improve phenotyping efforts, resulting in better training populations and faster improvement in shattering in the breeding population. Moreover, selecting for earlier maturing genets in the breeding program could also indirectly select for higher yields due to increased spike weight and seed size traits (documented here) and floret site utilization (Altendorf et al., 2021a).

Threshability, another important domestication trait that has received selection priority in the breeding program, increased across cycles on average, but was not statistically significant in our analyses. This insignificance is likely attributable to the lack of improvement in threshability from Cycle 3 to Cycle 4. This decrease in threshability from Cycle 3 to Cycle 4 was unexpected and could be due to a number of factors. Genetic drift causing random fluctuation of threshability alleles is one potential explanation. Another more likely explanation could be GxE or unique environmental effects in the years Cycle 3 was phenotyped and how those data were subsequently used to train and predict Cycle 4 parents. The UMN breeding program did not incorporate GxE into their genomic selection models until 2020, which was after Cycle 3 selections were made. After incorporating GxE into genomic selection models, model predictions for threshability improved by 23% (Bajgain et al., 2020a). Similarly, in a study using genomic prediction for threshability in a barley breeding population, researchers found high predictive ability using a multi-trait model but no improvement in predictive ability using threshability loci as covariates in the model (Massman et al., 2023). Thus, these findings

emphasize both the utility and model-optimization nuance in using genomic selection for trait improvement in IWG.

Trends for increased seed size traits and thousand grain weight (TGW) were promising and reflective of direct selection for those traits beginning with the parents of UMN-C3. Seed area and width increased slightly each cycle, and although TGW increases were not statistically significant, trends for relative gain each cycle were also encouraging. Our observed genetic gain in TGW of 1.38% per cycle, compared to the reported genetic gain for TGW of 0.30% per year in annual wheat from 1968 to 2006 (Underdahl et al., 2008), is encouraging. Of note, seed area, specifically the width of seeds or improved circularity, has previously been reported to be more difficult to increase than seed length (Lee Dehaan, personal communication). Additionally, our findings indicated a low broad-sense heritability value of $H^2 = .08$ for seed width, aligning with the general impression of difficulty to improve this trait in IWG populations despite others reporting relatively high broad-sense heritability for seed size traits (Zhang et al., 2016, 2017a; Bajgain and Anderson, 2021). Thus, these findings for improvements in area and width are encouraging for improving seed circularity.

Furthermore, relative gain for 10-spike seed and spike weight decreased slightly each cycle; these findings were unexpected. In the breeding program, 10-spike yield component traits appear to be generally increasing with each cycle. One likely explanation for this phenomenon, aside from being an anomaly, is that 10-spike traits are not directly selected upon. Other yield component traits such as seed size and TGW receive higher selection priority in the breeding program and generally have higher heritability (Zhang et al., 2016, 2017a; Bajgain and Anderson, 2021). Moreover, yield calculated from 10-spike samples taken from individual spaced plants are likely not the best indicator of yield in the highly competitive sward setting; for

example, in hybrid maize, individual plant yield potential has not increased over the past several decades. Yield increases in maize are due to the greater ability of the plant to tolerate high plant density (Duvick, 2001). Additionally, different genetic loci control biomass yield in the sward vs. spaced plant setting in IWG (Mortenson et al., 2019), which draws attention to the need for further studies assessing grain yield in the sward vs. spaced plant setting.

While comparing seed size of IWG to annual wheat is not necessarily a fair comparison, it is nonetheless interesting. If the average wheat seed area is $17mm^2$, it would take 86 cycles of selection at the current rate of gain to achieve a population mean seed area of $17mm^2$ in IWG. That statistic is not encouraging, but comes with multiple caveats, the first of which is the fact that seed size is only one of many yield component traits. Floret site utilization, number of spikelets per spike, tillers per plant, and many other factors affect IWG grain yield per unit area (Altendorf et al., 2021). Moreover, the two years (2021 & 2022) and locations (St. Paul and Lamberton, MN) where this study was grown both experienced hotter than average temperatures and lower than average precipitation from May to August. In June of 2021, both locations' average temperatures were the highest of the 20-year average (Supplemental Table S2). These heat and limited precipitation conditions around flowering and grain filling in the key summer months of both growing seasons and locations likely also influenced the yield component trait values reported in this study, as heat stress can cause kernel abortion and suppression of grain maturation (Hays et al., 2007) and drought stress reduces grain number and seed size (Wang et al., 2017). In addition, the value of IWG as a perennial grain crop extends well beyond grain production. Incorporating IWG into a cropping system provides biomass forage yield, requires fewer grower inputs, and mitigates nutrient leaching into groundwater resources (Jungers et al., 2019) - benefits that also come with economic and civic incentives for a grower.

In summary, this study sheds light on the true realized genetic gain for domestication and agronomic traits in UMN's IWG breeding population. We found that progress for these traits is positive, but there are a few actions which would greatly improve the rate of gain. The most impactful way to improve rate of genetic gain would be to reduce the time per cycle. At UMN, one breeding cycle spanned 2 years for UMN-C1 to UMN-C5, but since UMN-C6, one breeding cycle is completed every year. Starting in 2017 with the onset of genomic selection in their breeding program, TLI has completed one cycle per year, and in recent years, TLI has applied speed breeding techniques to achieve two cycles per year - a feat that greatly improves rates of gain in the context of time (Watson et al., 2018; Crain et al., 2021a). The application and improvement of genomic selection models, speed breeding techniques, high-throughput phenotyping and phenomics, and gene-editing applications are key to increasing the rate of genetic gain in IWG and subsequent widespread adoption.

1.7 References

Altendorf, K.R., L.R. DeHaan, G.C. Heineck, X. Zhang, and J.A. Anderson. 2021a. Floret site utilization and reproductive tiller number are primary components of grain yield in intermediate wheatgrass spaced plants. Crop Science 61(2): 1073–1088. doi: 10.1002/csc2.20385.

Altendorf, K.R., L.R. DeHaan, S.R. Larson, and J.A. Anderson. 2021b. QTL for seed shattering and threshability in intermediate wheatgrass align closely with well-studied orthologs from wheat, barley, and rice. The Plant Genome 14(3): e20145. doi: 10.1002/TPG2.20145.

Bajgain, P., and J.A. Anderson. 2021. Multi-Allelic Haplotype-Based Association Analysis Identifies Genomic Regions Controlling Domestication Traits in Intermediate Wheatgrass. Agriculture 2021, Vol. 11, Page 667 11(7): 667. doi: 10.3390/AGRICULTURE11070667.

Bajgain, P., J.L. Crain, D.J. Cattani, S.R. Larson, K.R. Altendorf, et al. 2022a. Breeding Intermediate Wheatgrass for Grain Production. Plant Breeding Reviews. John Wiley & Sons, Ltd. p. 119–217

Bajgain, P., C. Li, and J.A. Anderson. 2022b. Genome-wide association mapping and genomic prediction for kernel color traits in intermediate wheatgrass (Thinopyrum intermedium). BMC Plant Biol 22(1): 218. doi: 10.1186/s12870-022-03616-7.

Bajgain, P., X. Zhang, and J.A. Anderson. 2019. Genome-wide association study of yield component traits in intermediate wheatgrass and implications in genomic selection and breeding. G3: Genes, Genomes, Genetics 9(8): 2429–2439. doi: 10.1534/g3.119.400073.

Bajgain, P., X. Zhang, and J.A. Anderson. 2020a. Dominance and G×E interaction effects improve genomic prediction and genetic gain in intermediate wheatgrass (Thinopyrum intermedium). The Plant Genome 13(1): e20012. doi: 10.1002/tpg2.20012.

Bajgain, P., X. Zhang, J.M. Jungers, L.R. DeHaan, B. Heim, et al. 2020b. 'MN-Clearwater', the first food-grade intermediate wheatgrass (Kernza perennial grain) cultivar. Journal of Plant Registrations 14(3): 288–297. doi: 10.1002/PLR2.20042.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67: 1–48. doi: 10.18637/jss.v067.i01.

Cochran, W., and G. Cox. 1957. Experimental designs. 2nd ed. John Wiley & Sons, Inc.

Crain, J., P. Bajgain, J. Anderson, X. Zhang, L. DeHaan, et al. 2020. Enhancing Crop Domestication Through Genomic Selection, a Case Study of Intermediate Wheatgrass. Frontiers in Plant Science 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.00319 (accessed 15 May 2023).

Crain, J., L. DeHaan, J. Poland, C. Jesse Poland, and W. Genet-. 2021a. Genomic prediction enables rapid selection of high-performing genets in an intermediate wheatgrass breeding program. The Plant Genome 14(2): e20080. doi: 10.1002/TPG2.20080.

Crain, J., A. Haghighattalab, L. DeHaan, and J. Poland. 2021b. Development of whole-genome prediction models to increase the rate of genetic gain in intermediate wheatgrass (Thinopyrum intermedium) breeding. The Plant Genome 14(2): e20089. doi: 10.1002/TPG2.20089.

Crain, J., S. Larson, K. Dorn, L. DeHaan, and J. Poland. 2022. Genetic architecture and QTL selection response for Kernza perennial grain domestication traits. Theor Appl Genet 135(8): 2769–2784. doi: 10.1007/s00122-022-04148-2.

DeHaan, L., M. Christians, J. Crain, and J. Poland. 2018. Development and Evolution of an Intermediate Wheatgrass Domestication Program. Sustainability 10(5): 1499. doi: 10.3390/su10051499.
DeHaan, L.R., and B.P. Ismail. 2017. Perennial cereals provide ecosystem benefits. Cereal Foods World 62(6): 278–281. doi: 10.1094/CFW-62-6-0278.
DeHaan, L., S. Wang, S. Larson, D. Cattani, X. Zhang, et al. 2014. Current Efforts to Develop Perennial Wheat and Domesticate Thinopyrum intermedium as a Perennial Grain.
Donmez, E., R.G. Sears, J.P. Shroyer, and G.M. Paulsen. 2001. Genetic Gain in Yield Attributes of Winter Wheat in the Great Plains. Crop Science 41(5): 1412–1419. doi: 10.2135/cropsci2001.4151412x.
Duvick, D.N. 2005. Genetic Progress in Yield of United States Maize (Zea mays L.). Maydica 50: 193–202.
Eberhart, S.A. 1964. LEAST SQUARES METHOD FOR COMPARING PROGRESS AMONG RECURRENT SELECTION METHODS. Crop Science 4(2): 230–231.
Fedak, G., and F. Han. 2005. Characterization of derivatives from wheat-Thinopyrum wide crosses. Cytogenetic and Genome Research 109(1–3): 360–367. doi: 10.1159/000082420.
Hays, D.B., J.H. Do, R.E. Mason, G. Morgan, and S.A. Finlayson. 2007. Heat stress induced ethylene production in developing wheat grains induces kernel abortion and increased maturation in a susceptible cultivar. Plant Sci. 172(6): 1113–1123. doi: 10.1016/j.plantsci.2007.03.004.
Jungers, J.M., L.H. DeHaan, D.J. Mulla, C.C. Sheaffer, and D.L. Wyse. 2019. Reduced nitrate leaching in a perennial grain crop compared to maize in the Upper Midwest, USA. Agriculture, Ecosystems & Environment 272: 63–73. doi: 10.1016/J.AGEE.2018.11.007.
Kaye, J.P., and M. Quemada. 2017. Using cover crops to mitigate and adapt to climate change. A review. Agron. Sustain. Dev. 37(1): 4. doi: 10.1007/s13593-016-0410-x.
Massman, J.M., H.-J.G. Jung, and R. Bernardo. 2013. Genomewide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize. Crop Science 53(1): 58–66. doi: 10.2135/cropsci2012.02.0112.
Massman, C., B. Meints, J. Hernandez, K. Kunze, K.P. Smith, et al. 2023. Genomic prediction of threshability in naked barley. Crop Science 63(2): 674–689. doi: 10.1002/csc2.20907.
McLean, D., L. Apiolaza, M. Paget, and J. Klápště. 2023. Simulating deployment of genetic gain in a radiata pine breeding program with genomic selection. Tree Genetics & Genomes 19(4): 33. doi: 10.1007/s11295-023-01607-9.
Mortenson, J.S., B.L. Waldron, S.R. Larson, K.B. Jensen, L.R. DeHaan, et al. 2019. Quantitative Trait Loci (QTL) for Forage Traits in Intermediate Wheatgrass When Grown as Spaced-Plants versus Monoculture and Polyculture Swards. Agronomy 9(10): 580. doi: 10.3390/agronomy9100580.
Pimentel, D., D. Cerasale, R.C. Stanley, R. Perlman, E.M. Newman, et al. 2012. Annual vs. perennial grain production. Agriculture, Ecosystems and Environment 161: 1–9. doi: 10.1016/j.agee.2012.05.025.
Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2). doi: 10.1371/journal.pone.0032253.
R Core Team. 2021. R: A language and environment for statistical computing. https://www.R-project.org/.

Reilly, E.C., A. Conway-Anderson, J.G. Franco, J.M. Jungers, E.B. Moore, et al. 2023. Editorial: Continuous living cover: adaptive strategies for putting regenerative agriculture into practice. Frontiers in Sustainable Food Systems 7. https://www.frontiersin.org/articles/10.3389/fsufs.2023.1320870 (accessed 5 December 2023).

Rutkoski, J.E. 2019. Estimation of Realized Rates of Genetic Gain and Indicators for Breeding Program Assessment. Crop Science 0(0): 0. doi: 10.2135/cropsci2018.09.0537.

Scott, B.A., M. Haile-Mariam, B.G. Cocks, and J.E. Pryce. 2021. How genomic selection has increased rates of genetic gain and inbreeding in the Australian national herd, genomic information nucleus, and bulls. Journal of Dairy Science 104(11): 11832–11849. doi: 10.3168/jds.2021-20326.

Sharma, R.C., J. Crossa, G. Velu, J. Huerta-Espino, M. Vargas, et al. 2012. Genetic Gains for Grain Yield in CIMMYT Spring Bread Wheat across International Environments. Crop Science 52(4): 1522–1533. doi: 10.2135/cropsci2011.12.0634.

Sharma, H., H. Ohm, L. Goulart, R. Lister, R. Appels, et al. 1995. Introgression and characterization of barley yellow dwarf virus resistance from Thinopyrum intermedium into wheat. Genome 38(2): 406–413. doi: 10.1139/g95-052.

Tessema, B.B., H. Liu, A.C. Sørensen, J.R. Andersen, and J. Jensen. 2020. Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat. Front. Genet. 11. doi: 10.3389/fgene.2020.578123.

Underdahl, J.L., M. Mergoum, Joel.K. Ransom, and B.G. Schatz. 2008. Agronomic Traits Improvement and Associations in Hard Red Spring Wheat Cultivars Released in North Dakota from 1968 to 2006. Crop Science 48(1): 158–166. doi: 10.2135/cropsci2007.01.0018.

Vogel, K.P., K. Arumuganathan, and K.B. Jensen. 1999. Nuclear DNA Content of Perennial Grasses of the Triticeae. Crop Science 39(3): cropsci1999.0011183X003900020009x. doi: 10.2135/cropsci1999.0011183X003900020009x.

Vogel, K., and J. Jensen. 2001. Adaptation of perennial triticeae to the eastern Central Great Plains. Journal of Range Management 54: 674–679.

Wagoner, P. 1990. Perennial grain new use for intermediate wheatgrass. Journal of Soil and Water Conservation 45(1): 81–82.

Wang, J.-Y., Y.-C. Xiong, F.-M. Li, K.H.M. Siddique, and N.C. Turner. 2017. Effects of Drought Stress on Morphophysiological Traits, Biochemical Characteristics, Yield, and Yield Components in Different Ploidy Wheat: A Meta-Analysis. In: Sparks, D.L., editor, ADVANCES IN AGRONOMY, VOL 143. Elsevier Academic Press Inc, San Diego. p. 139–173

Watson, A., S. Ghosh, M.J. Williams, W.S. Cuddy, J. Simmonds, et al. 2018. Speed breeding is a powerful tool to accelerate crop research and breeding. Nature Plants 4(1): 23–29. doi: 10.1038/s41477-017-0083-8.

Zhang, X., S.R. Larson, L. Gao, S.L. Teh, L.R. DeHaan, et al. 2017a. Uncovering the Genetic Architecture of Seed Weight and Size in Intermediate Wheatgrass through Linkage and Association Mapping. The Plant Genome 10(3): plantgenome2017.03.0022. doi: 10.3835/plantgenome2017.03.0022.

Zhang, X., P. Pérez-Rodríguez, J. Burgueño, M. Olsen, E. Buckler, et al. 2017b. Rapid Cycling Genomic Selection in a Multiparental Tropical Maize Population. G3 Genes|Genomes|Genetics 7(7): 2315–2326. doi: 10.1534/g3.117.043141.

Zhang, X., A. Sallam, L. Gao, T. Kantarski, J. Poland, et al. 2016. Establishment and Optimization of Genomic Selection to Accelerate the Domestication and Improvement of Intermediate Wheatgrass. The Plant Genome 9(1): plantgenome2015.07.0059. doi: 10.3835/plantgenome2015.07.0059.

2. Genome-wide Association Study Identifies Genomic Regions under Selection for Domestication and Agronomic Traits in Intermediate Wheatgrass

2.1 Abstract

Intermediate wheatgrass (*Thinopyrum intermedium*, IWG) is a new perennial grain crop undergoing direct domestication for several traits. IWG as a perennial grain crop provides numerous ecosystem services and has the potential to benefit rural communities by providing an alternative crop option with a high market value. However, IWG has only been under development as a perennial grain crop for the past few decades, and at the University of Minnesota (UMN) since 2011. Thus, improvements in grain-related traits, market development, and establishment of best agronomic practices are key to the long-term viability of IWG as a Midwestern grain crop. In this study, 224 parent genets (genetically unique individuals) from UMN IWG breeding cycles 2, 3, 4, and 5 were cloned, planted, and evaluated over 2 years (2021 & 2022) in 2 locations (St. Paul & Lamberton, MN). Plants were genotyped using genotyping-by-sequencing to get single nucleotide polymorphisms (SNPs), and phenotyped for key domestication traits including shattering, brittle rachis, seed size traits, and spike characteristics. A genome-wide association study (GWAS) identified 33 quantitative trail loci (QTL) for shattering, seed size, and yield traits, which individually explained 13% of phenotypic variation in the population, on average. Some of the identified QTL were also identified in previous IWG studies, and some were novel, but appeared to be in close proximity to previously identified domestication traits. For example, a brittle rachis QTL on the J03 IWG chromosome. Moreover, changes in allele frequencies across breeding cycles for significant QTL were examined to identify any alleles which appeared to be under positive or negative selection. The frequency of

favorable alleles increased for several QTL/trait combinations (average shattering, spikelet density, thousand grain weight, seed length, etc.), and decreased for other QTL/trait combinations. These findings may be used to inform IWG breeding programs of key regions of the genome which should be targeted for selection improvements.

2.2 Introduction

In the past few decades, an agricultural paradigm shift from annual to more sustainable perennial systems has been conceptualized and proposed by many (Jackson, 1980; Chapman et al., 2022; Streit-Krug et al., 2023). Agricultural cropland occupies approximately 12-13% of the global land area; of the cropland worldwide, annual crop species occupy 90% and only 10% is occupied by perennial crops (fao.org, 2020). The majority of these annual cropping systems are made up of oilseeds, legumes, and cereal grains (Cox et al., 2006). This annual, monocultural agricultural system poses multiple environmental and socio-economic issues such as soil degradation and erosion, nutrient loss, loss of biodiversity, dependency on outside inputs, and more - issues that could be mitigated with the adoption of a perennial grain crop system (Crews et al., 2018). This poses the question - can annual oilseed, legume, and cereal grain crops effectively be replaced with perennial alternatives that offer more ecosystem benefits?

In the first wave of agricultural domestication, herbaceous annuals were selected by humans and eventually made up what we now know as annual grain crops. Non-woody, herbaceous perennials were not domesticated, most likely due to perennial crops' tendency to have lower annual sexual reproductive output (Van Tassel et al., 2010). Several have suggested strategies to effectively domesticate perennial grain crops in the modern age (Cox et al., 2002;

DeHaan et al., 2016, 2020; Crews and Cattani, 2018). In the dawn of affordable phenotyping and genotyping technologies, rapid domestication of new crops is more feasible.

For perennial grain crops, specifically intermediate wheatgrass (IWG, *Thinopyrum intermedium* (Host) Barkworth & D.R. Dewey), to be successfully commercialized for widespread use as a perennial grain crop, several plant breeding and agronomic characteristics must be improved. IWG, a perennial forage grass, was originally selected as a promising candidate for perennial grain domestication by scientists at the Rodale Institute (Kutztown, PA, USA) in the late 1980s due to its larger seed size and spikelet fertility (Wagoner, 1990; Wagoner and Schaeffer, 1990). Since then, multiple IWG breeding programs across the US and beyond have been initiated to improve grain yield, seed size, and other key domestication traits. Notably, The Land Institute (TLI) began breeding efforts in 2001 - selecting for a large panel of traits including reduced shattering, improved free-threshing, seed yield, and more (DeHaan et al., 2018). In 2011, the University of Minnesota (UMN) received 2560 seedlings from 66 Cycle 3 families from TLI's IWG breeding program, which were used to initiate an IWG breeding program in Minnesota (Zhang et al., 2016). At the time of this publication, the UMN IWG Breeding program has completed 7 selection cycles.

IWG has a large (12.75 Gb), allohexaploid (2n = 6x = 42) genome, and populations are highly allelically diverse due to its obligate outcrossing nature (Vogel et al., 1999; Bajgain et al., 2022a). A high-quality reference genome is available through the US Department of Energy Joint Genome Institute (Thinopyrum intermedium Genome Sequencing Consortium). Previous studies have leveraged genomic resources to identify quantitative trait loci (QTL) for various traits in IWG, many of which have found these QTL align well with orthologous genes in other domesticated species. For example, IWG QTL for the timing of anthesis align with barley

flowering-timing genes (Altendorf et al., 2021b), various domestication trait QTL align with orthogenes for the wheat *Q* and *RHT* genes (Larson et al., 2019), a brittle rachis shattering QTL was found in close proximity to a barley *Btr2* gene ortholog (Crain et al., 2022), and several QTL for free-threshing and shattering align closely with known domestication genes such as *SH5* and *SHAT1* in rice, *Btr2*, *Btr1*, and *Btr1*-like in barley, and *Q*, *sog*, and *Tg* in wheat (Altendorf et al., 2021a).

Across previous IWG studies, genetic control for both domestication and agronomic traits appears to be under relatively quantitative, polygenic control. Crain et al. found domestication traits to be controlled by 33 to 558 different QTL, each accounting for less than 5% of phenotypic variance, depending on the trait (2022). Moreover, Altendorf et al. found brittle rachis, threshability, and particularly floret shattering, are controlled by multiple loci with relatively small effects (2021a). These findings draw attention to the applicability and utility of genomic selection (GS) in IWG breeding programs. GS leverages molecular markers paired with phenotypic data to create a training set that predicts the performance of other unphenotyped individuals (Meuwissen et al., 2001). With the dawn of affordable marker technologies, GS is an excellent tool to reduce costs and increase genetic gain per unit of time (Crossa et al., 2017). Thus, the incorporation of GS into IWG breeding programs is now both routine and successful at achieving acceptable prediction accuracies for both domestication and agronomic traits (Zhang et al., 2016; Crain et al., 2020). In addition, many have leveraged QTL identified through genome-wide association studies (GWAS) to improve GS models. GWAS, or association mapping, is a widely used method to discover QTL in diverse populations using genome-wide markers in association with traits of interest (Weir, 2008). In IWG, researchers have used identified QTL for disease resistance (Bajgain et al., 2019b), kernel color (Bajgain et al., 2022b), and other yield

component and domestication traits (Zhang et al., 2017; Bajgain et al., 2019a; Bajgain and Anderson, 2021) to improve the accuracy of GS models as well.

IWG genetic mapping research at the University of Minnesota has previously focused on a curated biparental population (Altendorf et al., 2021a), or a snapshot of one breeding cycle as a population (Bajgain et al., 2019a). This research utilizes parents from four different UMN breeding cycles, which better captures the breadth of allelic diversity contained in the UMN IWG breeding population. Moreover, previous research has demonstrated meaningful gains in phenotypic values for domestication and agronomic traits in the UMN IWG breeding program (Stoll et al., 2024, in press), however, insights into how phenotypic progress for key traits relates to changes in allelic frequency for key domestication and agronomic traits has not been investigated in IWG. Thus, the objectives of this research are to 1) identify genomic regions (QTL) associated with various agronomic and domestication traits and 2) assess how allele frequencies for said traits change across breeding cycles. This research will provide insights into genomic regions that have been directly or indirectly selected as breeding for agronomic and domestication traits has progressed, and allow scientists to make recommendations for future breeding targets.

2.3 Materials and Methods

2.3.1 Plant Materials & Phenotyping

Plant materials for this study consisted of parent genets, genetically unique plants, from cycles 2, 3, 4, and 5 of the University of Minnesota's (UMN) intermediate wheatgrass (IWG) breeding program. Cycle 2, 3, and 4 genets were obtained in the 2020 summer by cloning plants from a living nursery of parent plants from each breeding cycle. Direct vegetative clones were necessary

for true replication as IWG is an obligate-outcrossing species with synthetic varieties. Cycle 5 parents were similarly cloned directly from the plants selected as parents for the Cycle 5 breeding population, initiated in 2020. In total, 242 genets were cloned and planted in an unreplicated trial at 2 locations: St. Paul, MN and Lamberton, MN, as described in Stoll et. al., 2024 (in press). In the final association analyses, 223 genets were included due to some plants dying after transplanting. Trials were planted in the fall of 2020 and phenotypic data were collected in the 2021 and 2022 growing seasons.

In 2021 and 2022, plants were phenotyped for height, anthesis date, 10-spike seed and spike dry weight, spikelet density, brittle rachis shattering, floret shattering, threshability, thousand grain weight (TGW), and seed size traits. To measure the preharvest traits height and anthesis date, the height in centimeters was recorded using a measuring stick in the center of the plant, and anthesis date in Julians (days after January 1st) was recorded when approximately half of the plant was actively shedding pollen. To phenotype post-harvest traits, 10 spikes were harvested from each plant when most of the experiment was at optimal harvest timing and carefully stored and dried to minimize shattering before phenotyping. Spikes were weighed to get 10-spike dry weight in grams, and a subsample of 6 spikes was used to calculate spikelet density by taking the spikelet count per spike divided by the spike length in centimeters. Then, two types of shattering were measured: brittle rachis shattering and floret shattering, as they are independently genetically controlled (Larson et al., 2019; DeHaan et al., 2020; Altendorf et al., 2021a). Shattering methods were adapted from methods reported by DeHaan et al. (2018), and Altendorf et al. (2021a).To measure shattering, the subsample of 6 spikes was dropped one by one from a consistent one-meter height onto a metal lab bench three times and then rated for the two types of shattering on a 0 to 9 scale. For floret shattering, ratings ranged from 0 denoting no

shattering (0% shattering) to 9 indicating extensive shattering (91 - 100% shattering), with intermediate ratings reflecting varying degrees of shattering intensity. Likewise, brittle rachis shattering ratings ranged from 0 representing no rachis breaks to 9 indicating nine or more rachis breaks, with intermediate ratings capturing incremental levels of rachis breakage.

Samples were threshed using a Wintersteiger LD 350 (Wintersteiger Inc, Salt Lake City, USA) using settings that would reduce seed breakage and blower settings to keep hulled and lighter seeds in the sample. Then, the full sample's weight in grams was recorded. Next, samples were assessed for threshability, indicating the portion of hulled versus naked seeds using a visual rating scale ranging from 0 to 9. A rating of 0 denoted 0-10% of the sample were naked seeds, while a rating of 9 indicated 91-100% of seeds were naked. Next, 50-150 naked seeds were subsampled to record thousand grain weight (TGW) in grams, seed area, length, and width in mm^2, and seed length-to-width ratio using the Marvin Optical Seed Analyzer (GTA Sensorik GmbH, Germany).

2.3.2 Phenotypic Data Analysis

We used a linear mixed-effects model to investigate the relationship between the trait of interest and genetic and environmental factors using the lme4 package with R software (Bates et al., 2015; R Core Team, 2021). BLUEs (Best Linear Unbiased Estimates) for each genet were calculated as follows for all traits:

$$Y_{ij} = \text{Environment}_i + \text{Genet}_j + \epsilon_{ij}$$

Where Y_{ij} is the observed value for trait at the i-th environment under the j-th genet, Environment_i denotes the random effect of the ith Environment, Genet_j denotes the fixed effect of the jth genet, and ϵ_{ij} is the random error term $N(0, \sigma^2)$.

Broad-sense heritability on the entry-mean basis was calculated as follows:

$$H^2 = \boldsymbol{\sigma}_G^2 / (\boldsymbol{\sigma}_G^2 + \boldsymbol{\sigma}_{Error}^2 / E)$$

Where $\boldsymbol{\sigma}_G^2$ is the genetic variance, $\boldsymbol{\sigma}_{Error}^2$ is the residual error variance, and E is the number of individual environments (each year by location combination is considered a different environment). The statistical model used to gather variance components for heritability calculations was fit as follows:

$$Y_{ij} = \text{Genet}_i + \text{Environment}_j + \epsilon_{ij}$$

Where Y_{ij} is the observed trait value the *i*-th genet under the *j*-th environment, Genet_i denotes the random effect of the *i*th genet, Environment_j denotes the fixed effect of the *j*th environment, and ϵ_{ij} is the random error term $N(0, \boldsymbol{\sigma}^2)$.

2.3.3 Genotyping, SNP Calling Pipeline, & Imputation

Leaf tissue from each genet at each location was harvested and dried, and DNA was extracted using the BioSprint 96 Plant DNA Kit (QIAGEN) and normalized to 10 ng/µL. Genotyping by sequencing libraries were developed using the *Pst*I and *Msp*I enzymes to digest samples and coded with two barcodes per sample. Each 96-well plate was pooled for a total of six libraries that were then sequenced using genotyping by sequencing (Elshire et al., 2011; Poland et al., 2012) on the Illumina Novaseq 6000. After de-multiplexing, single nucleotide polymorphisms (SNPs) for each genet were aligned to v3.1 of the Thinopyrum intermedium reference genome assembly (Thinopyrum intermedium Genome Sequencing Consortium) using bwa (Li and Durbin, 2009), indexed using SAMtools, and variants were called using BCFtools (Danecek et al., 2021). Hard quality filters for RMS mapping quality greater than or equal to 30 and combined depth across samples (INFO/DP) greater than or equal to 5 were applied. Further

filters were applied to select for minimum read depth greater than 3, only biallelic SNPs, genets with less than 95% missingness, variants with less than 80% missing genotypes, and minor allele frequency greater than 0.03. Imputation was performed using the LD-kNNi method using default settings (Money et al., 2015) in Tassel version 5.2.93 (Bradbury et al., 2007).

2.3.4 Linkage Disequilibrium

Linkage disequilibrium (LD) was calculated in Tassel version 5.2.93 (Bradbury et al., 2007) across the whole genome using a sliding window of 1000 markers. Additionally, LD was calculated for each chromosome using all pairwise comparisons within a given chromosome. LD decay was estimated using the Hill and Weir formula (Hill and Weir, 1988) with the *nls* function in R. The half-decay distance was considered the threshold of LD decay (Flint-Garcia et al., 2003).

2.3.5 Genome-Wide Association Mapping

Genome-wide association (GWA) mapping was accomplished using the FarmCPU model (Liu et al., 2016) within GAPIT Software, version 3.1.0 (Lipka et al., 2012). To account for familial relatedness, a kinship matrix was calculated using TASSEL software default settings (version 5.2.93) and included in the model (Bradbury et al., 2007). Significant SNPs were detected at the Bonferroni adjusted p-value of $\alpha = 0.05$, divided by the ~12K SNPs in the dataset ($p < 4.14E\text{-}6$). Percent (phenotypic) variance explained (PVE) by each SNP in the population was calculated using the formula from Broman & Sen (2009, p. 246).

2.3.6 Allele Frequency Changes Across Breeding Cycles

To identify evidence of selection for several traits in this population of IWG breeding parents from multiple cycles, allele frequencies of significant SNPs detected in GWAS from each breeding cycle were calculated. In addition, the favorable allele, or the allele conferring a desirable phenotypic response in the homozygous and heterozygous states, was identified for each SNP by taking the mean phenotypic value for each genotype for each SNP (Table 2.3). Major, minor, and favorable allele frequencies and the proportion of homozygous major and minor individuals were calculated for each SNP and breeding cycle. Trends in favorable allele frequency across the breeding cycles were plotted for each SNP.

2.3.7 Comparison of Identified QTL to Domesticated Species and Other IWG Studies

FASTA sequences for significant genomic regions detected in association mapping were attained by first resolving each significant SNP to a quantitative trait locus (QTL) by adding the LD half-decay distance (.78 Mbp) to the left and right of the SNP. Then, this positional information was compared to the IWG v3.1 annotation file (Thinopyrum intermedium Genome Sequencing Consortium) to get both nucleotide and protein FASTA files using SAMtools (Li et al., 2009). Then, FASTA protein sequences were compared against several other species using BLASTP (Altschul et al., 1990) to identify putative homologous proteins from the related domesticated species wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), and barley (*Hordeum vulgare* L.). The top 10 identified proteins with the highest percent identity to IWG protein sequences are contained in Supplemental Table S5. Additionally, a literature search was conducted to compare identified IWG QTL with those found in other recent IWG studies.

2.4 Results

2.4.1 SNP Discovery Pipeline & Linkage Disequilibrium

After filtering and imputation, 12,072 SNPs remained, with an average of 574 SNPs per chromosome and a range of 389 SNPs (V01) to 811 SNPs (J07) on any given chromosome. Linkage disequilibrium (LD) and LD decay was calculated across each chromosome and genome-wide. Overall, LD in the IWG population decayed rapidly. Across individual chromosomes, LD, as determined by the half-decay distance, for some chromosomes was less than 600 bp (J04, J06, J07, S02, S03, S06, V02, V03, and V04), whereas chromosome V07 had the highest LD of 1.09Mbp (Supplemental Table S3). LD decay calculated genome-wide was estimated to be .78 Mbp. LD decay of .78 Mbp was used to resolve QTL for this study.

2.4.2 Brittle Rachis, Floret, and Average Shattering

Heritability in the broad sense (H^2) was high for brittle rachis ($H^2 = 0.80$), floret shattering ($H^2 = 0.81$), and average shattering ($H^2 = 0.79$, Table 2.1. In general, phenotypic values for shattering were slightly higher (i.e. worse shattering) for both types of shattering (and average shattering) at St. Paul than in Lamberton in both years (Table 2.1). Brittle rachis and floret shattering were both positively correlated with 10-spike seed weight ($r = 0.34, 0.22$, Figure 2.1) and floret shattering was positively correlated with TGW, seed area, width, and length ($r = 0.29, 0.28, 0.25, 0.20$, Figure 2.1). Four brittle rachis QTL were found on chromosomes J03, S02, S03, and S07, accounting for 11.9% to 16.3% (J03) of phenotypic variation (Table 1). One QTL was found for floret shattering on chromosome V05 accounting for 14.7% of the phenotypic variation, and four QTL were found for average shattering on chromosomes J02,

S01, S06, and V03, which accounted for 12.0% to 18.3% (J02) of phenotypic variation (Table 2.2).

Allele frequencies for identified QTL varied by individuals with each breeding cycle; some SNPs appeared to be under selection for or against the favorable allele with each successive breeding cycle. The favorable allele, or the allele that confers a positive phenotypic response in the homozygous and heterozygous states, was identified for each SNP (Table 2.2). For brittle rachis, three loci appeared to be under slight selection against the favorable allele (J03, S02, and S07), and one locus (S03) does not exhibit any dramatic evidence of selection (favorable allele frequency increase of 0.030 from Cycle 2 to 5, Supplemental Table S4). For the brittle rachis loci on J03, S07, and S02, the favorable allele frequency from Cycle 2 to Cycle 5 decreased by 0.062 (J03), 0.040 (S02), and by 0.046 (S07, Figure 2.2, Supplemental Table S4). For the single locus identified for floret shattering on chromosome V05, the favorable (minor) allele appeared to be under negative selection; the favorable allele frequency decreased by 0.131 from Cycle 2 to Cycle 5 (Figure 2.2, Supplemental Table S4). Two of the four loci identified for average shattering appeared to be under positive selection for the favorable (major) allele. The favorable allele frequency of the average shattering J02 locus increased by 0.090 from Cycle 2 to 5; likewise, the favorable allele frequency of the S06 locus increased by 0.035 from Cycle 2 to 5 (Figure 2.2, Supplemental Table S4). Additional information on identified proteins from BLASTP analyses with related domestication species can be found in Supplemental Table S5.

2.4.3 Seed Size

Heritability in the broad sense (H^2) was high for seed size traits, ranging from $H^2 = 0.78$ for seed width up to 0.88 for Seed L:W Ratio (Table 2.1). As mentioned in the previous section,

seed size traits were positively correlated with shattering. One QTL was found on chromosome V01 for seed width, accounting for 11.40% of phenotypic variation (Table 2.2). Nine QTL were found for seed length on chromosomes J03, J04, J05, S01, S03, S05, V02, and V07, accounting for 10.57% to 15.71% of phenotypic variation (Table 2.2). Notably, the seed width QTL found on J04 was also found significant for seed area in the GWAS analysis. The QTL for seed width on chromosome V01 appeared to be under positive selection, with favorable allele frequency increasing by 0.143 from Cycle 2 to 5 (Figure 2.2, Supplemental Table S4). Of the 9 QTL for seed length, 5 appeared to be neutral (no change in allele frequency) or under selection for the favorable allele (Figure 2.2). The QTL on J04 for seed area and length appeared to be under negative selection, with favorable allele frequency decreasing by 0.106 from Cycle 2 to Cycle 5 (Figure 2.2, Supplemental Table S4).

2.4.4 TGW, Spikelet Density, & 10-Spike Seed and Spike Weight

Heritability in the broad sense (H^2) was high for spikelet and grain weight traits, ranging from $H^2 = 0.70$ for 10-spike weight up to 0.82 for spikelet density (Table 2.1). 10-spike weight and seed weight were positively correlated with spikelet number ($r = 0.43, 0.28$, Figure 2.1), and TGW was negatively correlated with spikelet number and density ($r = -0.22, -0.23$, Figure 2.1). Again, shattering traits were positively associated with spike and seed weight traits. Six QTL for spikelet density were found on chromosomes J07, S02, S05, S07, and V02, explaining 11.16% to 19.54% of phenotypic variation (Table 2.2). Three QTL for TGW were found on chromosomes J03, J06, and S03, explaining 10.76% to 17.75% of the variation (Table 2.2). Finally, 2 QTL were found for 10-spike seed weight on J05 and S05, and 3 QTL were found for 10-spike weight on J02, S05, and V02 (Table 2.2). Five of the six QTL for spikelet density appeared to be under

neutral or positive selection, and all three TGW QTL were under neutral or positive selection for the favorable allele (Figure 2.2, Supplemental Table S4). All 5 QTL for 10-spike seed and spike weight appeared to be under neutral or negative selection for the favorable allele (Figure 2.2, Supplemental Table S4).

2.5 Tables and Figures

Table 2.1: Mean and standard deviation calculated in each environment and broad-sense heritability calculated on an entry mean basis is displayed for each trait.

Trait	H^2	Environment	Mean	Standard Deviation
Height	0.703	Lamb_2021	117.79	16.58
		Lamb_2022	126.88	9.93
		StP_2021	145.40	16.54
		StP_2022	139.50	12.27
Anthesis	0.721	Lamb_2021	172.06	2.69
		Lamb_2022	178.87	1.20
		StP_2021	167.75	2.10
		StP_2022	175.48	1.46
Threshability	0.846	Lamb_2021	6.33	1.79
		Lamb_2022	4.42	2.06
		StP_2021	5.26	1.98
		StP_2022	3.69	2.01
Brittle Rachis	0.804	Lamb_2021	1.67	0.96
		Lamb_2022	1.65	0.90
		StP_2021	2.02	1.26
		StP_2022	2.13	1.16
Floret Shattering	0.811	Lamb_2021	2.53	1.76
		Lamb_2022	2.08	1.71
		StP_2021	3.13	1.77
		StP_2022	3.29	2.06
Average Shattering	0.792	Lamb_2021	2.10	1.04
		Lamb_2022	1.87	0.91
		StP_2021	2.57	1.05
		StP_2022	2.71	1.10
Spike Length	0.686	Lamb_2021	27.61	3.84
		Lamb_2022	25.22	2.67
		StP_2021	32.30	3.58
		StP_2022	25.62	2.82
Spikelet Number	0.72	Lamb_2021	21.35	2.59
		Lamb_2022	21.28	2.28
		StP_2021	24.17	2.28
		StP_2022	21.80	2.45
Spikelet Density	0.816	Lamb_2021	0.78	0.09
		Lamb_2022	0.85	0.10
		StP_2021	0.76	0.09
		StP_2022	0.86	0.11

10-Spike Weight	0.695	Lamb_2021	13.17	3.26
		Lamb_2022	9.22	1.85
		StP_2021	18.01	3.79
		StP_2022	10.37	2.27
10-spike Seed Weight	0.804	Lamb_2021	6.17	2.44
		Lamb_2022	4.62	1.49
		StP_2021	10.13	2.89
		StP_2022	5.37	1.86
TGW	0.768	Lamb_2021	8.82	1.62
		Lamb_2022	8.37	1.55
		StP_2021	10.68	1.98
		StP_2022	7.92	1.41
Seed Area	0.802	Lamb_2021	8.53	1.02
		Lamb_2022	8.42	0.94
		StP_2021	9.83	1.23
		StP_2022	8.46	0.96
Seed Width	0.788	Lamb_2021	1.79	0.12
		Lamb_2022	1.76	0.11
		StP_2021	1.92	0.14
		StP_2022	1.74	0.11
Seed Length	0.866	Lamb_2021	6.25	0.44
		Lamb_2022	6.30	0.45
		StP_2021	6.63	0.46
		StP_2022	6.42	0.45
Seed L:W Ratio	0.881	Lamb_2021	3.53	0.27
		Lamb_2022	3.63	0.27
		StP_2021	3.50	0.25
		StP_2022	3.74	0.28

Table 2.2: Significant SNPs found in GWA analysis for key traits. Chromosome, basepair position, major/minor alleles, favorable alleles, minor allele frequencies (MAF), -log(p) value, and percent phenotypic variance explained (PVE) are displayed.

Trait	SNP	Chromosome	Position	Major/Minor Allele	Favorable Allele	MAF	-log(p)	PVE
			bp					%
Brittle Rachis	SJ03_171120774	J03	171120774	A / C	C	0.052	8.63	16.32
Brittle Rachis	SS02_462148698	S02	462148698	C / G	C	0.286	6.15	11.93
Brittle Rachis	SS03_37982539	S03	37982539	A / C	A	0.248	6.26	12.14
Brittle Rachis	SS07_96621872	S07	96621872	A / G	G	0.041	6.56	12.68
Floret Shattering	SV05_565531444	V05	565531444	T / C	C	0.381	7.71	14.72
Average Shattering	SJ02_633868944	J02	633868944	T / G	T	0.214	9.80	18.32
Average Shattering	SS01_265896448	S01	265896448	C / T	T	0.318	6.26	12.12
Average Shattering	SS06_60087375	S06	60087375	G / A	G	0.113	6.12	11.87
Average Shattering	SV03_172019389	V03	172019389	T / C	T	0.293	6.20	12.02
Spikelet Density	SJ07_436189945	J07	436189945	G / A	G	0.047	8.19	15.56
Spikelet Density	SS02_373179566	S02	373179566	A / G	G	0.178	10.53	19.54
Spikelet Density	SS02_465837109	S02	465837109	T / G	G	0.250	5.73	11.16
Spikelet Density	SS05_30587500	S05	30587500	G / A	G	0.137	7.94	15.11
Spikelet Density	SS07_82857819	S07	82857819	G / A	G	0.074	7.99	15.20
Spikelet Density	SV02_457071045	V02	457071045	C / A	C	0.464	5.77	11.23
10-spike Weight	SJ02_133239962	J02	133239962	C / T	C	0.498	8.47	16.05
10-spike Weight	SS05_161687303	S05	161687303	C / T	C	0.230	5.86	11.39
10-spike Weight	SV02_81709239	V02	81709239	G / A	G	0.162	5.40	10.54

10-spike Seed Weight	SJ05_21359377	J05	21359377	A / T	A	0.221	5.89	11.45
10-spike Seed Weight	SS05_56085353	S05	56085353	G / C	C	0.365	6.23	12.07
Seed Area	SJ03_461795327	J03	461795327	C / T	C	0.090	7.91	15.07
Seed Area	SJ04_507603982	J04	507603982	C / T	T	0.311	7.30	13.99
Seed Length	SJ03_461795327	J03	461795327	C / T	C	0.090	5.68	11.07
Seed Length	SJ04_516190345	J04	516190345	A / C	A	0.083	6.88	13.24
Seed Length	SJ05_484868935	J05	484868935	T / C	T	0.489	5.41	10.57
Seed Length	SS01_375800564	S01	375800564	A / G	G	0.169	5.42	10.58
Seed Length	SS03_357352971	S03	357352971	G / A	A	0.140	6.70	12.92
Seed Length	SS05_323670052	S05	323670052	T / C	T	0.417	8.27	15.71
Seed Length	SV02_106373759	V02	106373759	G / A	A	0.187	6.74	13.00
Seed Length	SV07_486034824	V07	486034824	A / G	G	0.054	6.86	13.21
Seed Width	SV01_409991737	V01	409991737	G / A	G	0.230	5.86	11.40
TGW	SJ03_28234197	J03	28234197	G / A	G	0.430	5.52	10.76
TGW	SJ06_540710206	J06	540710206	C / T	C	0.475	5.68	11.07
TGW	SS03_362322005	S03	362322005	T / C	C	0.151	9.46	17.75

Figure 2.1: Pearson correlation plot of all traits found significant at the r = 0.001***, 0.01**, and 0.05* levels. Insignificant correlations are denoted by "ns."

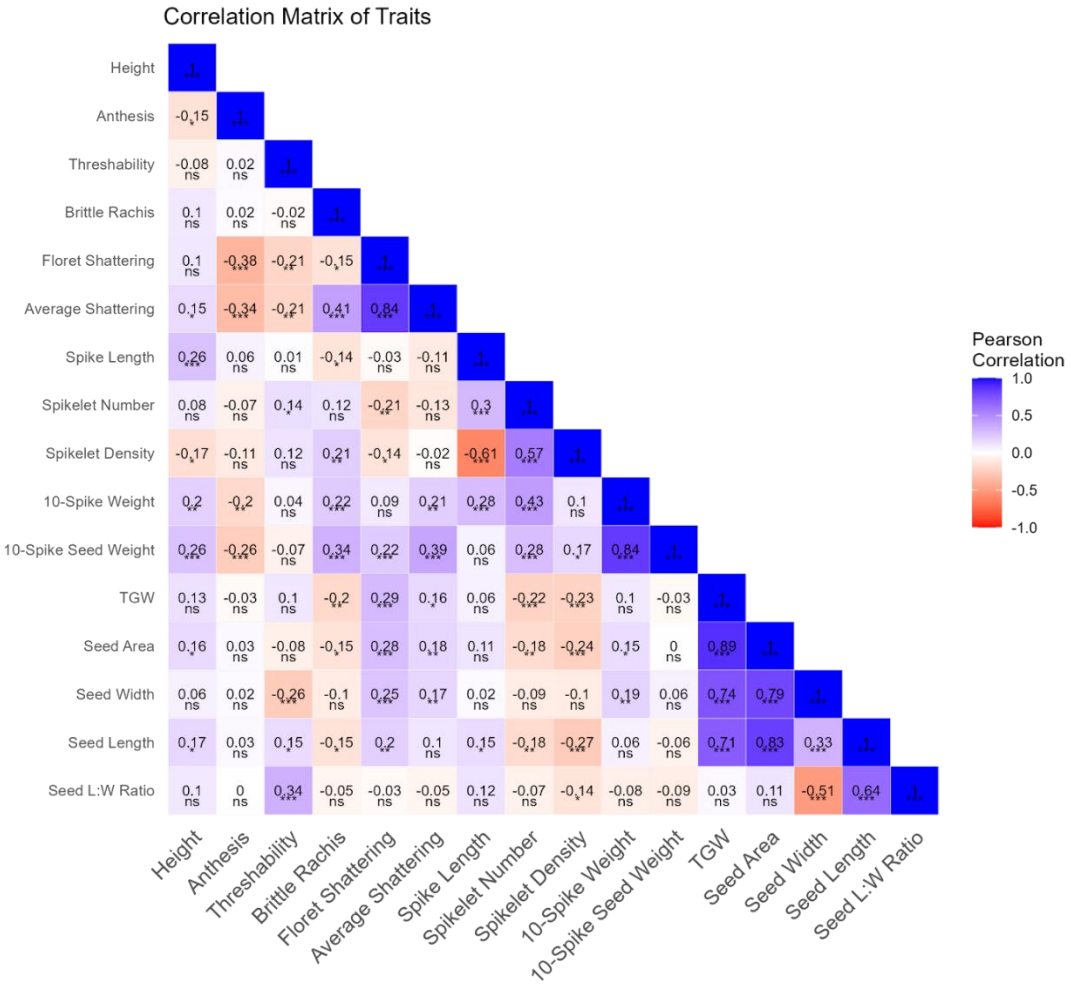


Figure 2.2: Favorable Allele Frequencies of SNPs found significant in GWAS, separated by breeding cycles.

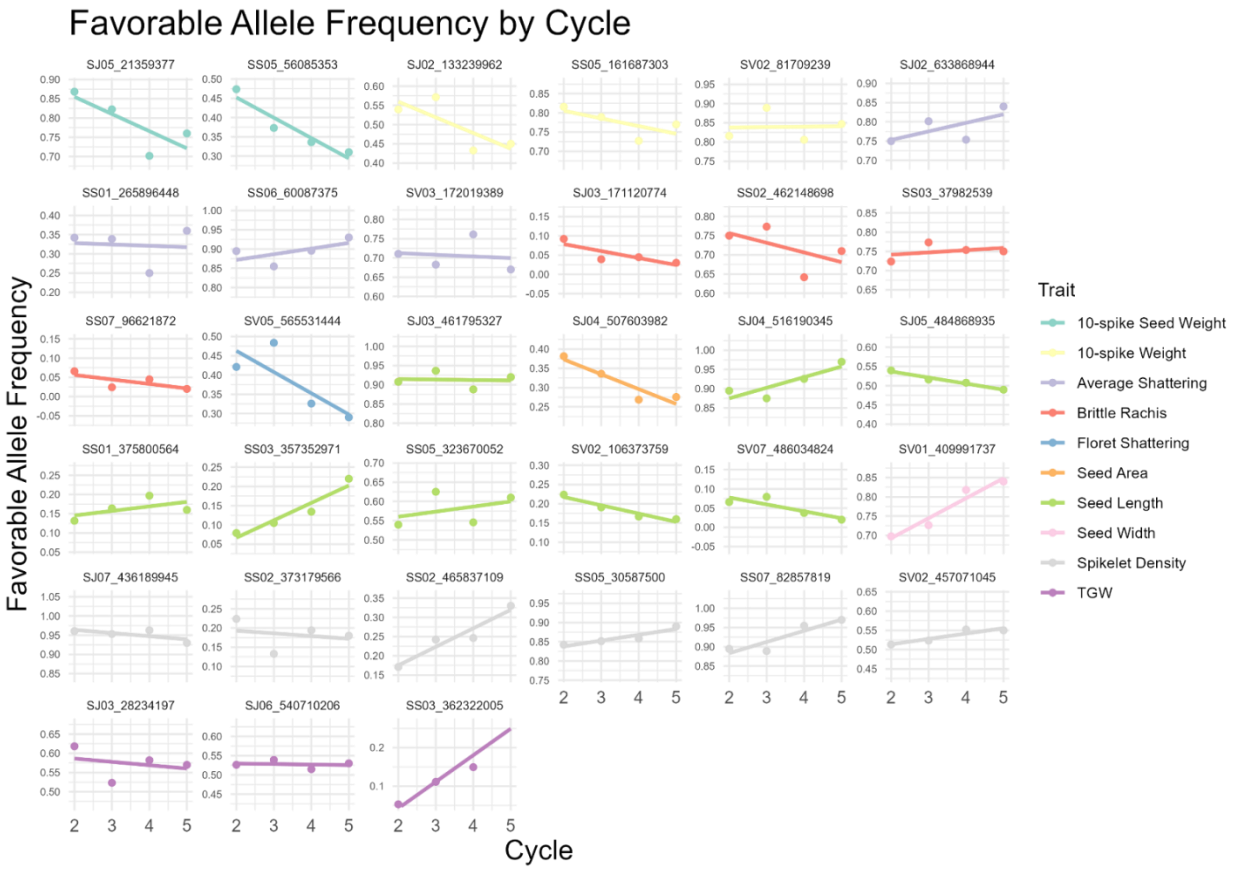


Table 2.3: Genotypes denoted by "A", "T", "C", or "G" are homozygous for the respective nitrogenous base. "R" denotes “A:G” heterozygote, "Y" denotes “C:T” heterozygote, "S" denotes “C:G” heterozygote, "W" denotes “A:T” heterozygote, "K" denotes “G:T” heterozygote, and "M" denotes “A:C” heterozygote.

Trait	SNP	Genotype*	Trait Mean
Brittle Rachis	SJ03_171120774	A	2.13
		C	0.21
		M	1.19
	SS02_462148698	C	1.88
		G	2.33
		S	2.27
	SS03_37982539	A	1.77
		C	2.70
		M	2.35
	SS07_96621872	A	2.11
		R	1.17
Floret Shattering	SV05_565531444	C	2.30
		T	3.52
		Y	2.98
Average Shattering	SJ02_633868944	G	3.52
		K	2.79
		T	2.40
	SS01_265896448	C	2.93
		T	2.38
		Y	2.34
	SS06_60087375	A	3.61
		G	2.38
		R	3.21
	SV03_172019389	C	3.29
		T	2.28
		Y	2.77
Spikelet Density	SJ07_436189945	G	0.82
		R	0.71
	SS02_373179566	A	0.79
		R	0.84
	SS02_465837109	G	0.80
		K	0.84

		T	0.78
	SS05_30587500	A	0.68
		G	0.82
		R	0.78
	SS07_82857819	G	0.82
		R	0.73
	SV02_457071045	A	0.72
		C	0.86
		M	0.80
10-spike Weight	SJ02_133239962	C	14.71
		T	13.15
		Y	14.07
	SS05_161687303	C	14.55
		T	13.59
		Y	13.24
	SV02_81709239	A	10.57
		G	14.43
		R	13.15
10-spike Seed Weight	SJ05_21359377	A	7.66
		T	6.73
		W	7.34
	SS05_56085353	C	8.35
		G	6.98
		S	7.71
Seed Area	SJ03_461795327	C	9.25
		Y	8.53
	SJ04_507603982	C	8.82
		Y	9.30
Seed Width	SV01_409991737	A	1.74
		G	1.86
		R	1.79
Seed Length	SJ04_516190345	A	6.58
		C	6.02
		M	6.24
	SJ05_484868935	C	6.38
		T	6.57
		Y	6.55
	SS01_375800564	A	6.40

		R	6.74
	SS03_357352971	A	6.97
		G	6.44
		R	6.73
	SS05_323670052	C	6.33
		T	6.75
		Y	6.42
	SV02_106373759	G	6.42
		R	6.69
	SV07_486034824	A	6.51
		R	6.61
TGW	SJ03_28234197	A	7.61
		G	9.79
		R	9.19
	SJ06_540710206	C	9.95
		T	9.19
		Y	9.21
	SS03_362322005	C	10.78
		T	8.97
		Y	9.98

2.6 Discussion

In this study, LD decayed rapidly across the genome (0.78 Mbp) and varied by chromosome, but overall was relatively low. These findings are consistent with previous studies identifying genome-wide LD around 0.375 Mbp (Crain et al., 2022) and 0.7 Mbp (Bajgain et al., 2019a). LD decay of distances less than 1 Mbp is relatively rapid, but not surprising due to the outcrossing nature of IWG allowing the crop to experience frequent recombination events.

A stringent Bonferroni-adjusted p-value of $p < 4.14E-6$ was used to detect QTL in the GWA study, leading to 33 QTL being discovered for all traits, one of which was common between seed area and length. Moreover, the relative PVE by the SNPs found in this study was relatively high

- ranging from 10.54% (10-spike weight) to 19.54% (spikelet density). One potential explanation for this finding is this study's relatively small population size (224). The number of QTL in a study is most accurately predicted with larger numbers of individuals and markers evaluated, rather than genetic variance and mean differences between parents (Dudley, 2007). Previous GWA studies in IWG have typically had over 1000 individuals in their populations (Bajgain et al., 2019a; Altendorf et al., 2021a; Crain et al., 2022). Despite calculating relatively high PVE for several QTL in this study, these traits are likely all under polygenic, quantitative control. Crain et al. (2022) calculated the theoretical minimum population size necessary to detect QTLs explaining 50% of the genotypic variance in IWG and found a minimum population size of 624 for brittle rachis, 2823 for shattering (what we call “floret shattering” in this study), 809 for free-threshing (threshability), and 93 for seed area. Furthermore, they calculated the estimated number of QTL controlling several traits and found the estimated QTL for brittle rachis to be 166, 187 for shattering, 304 for free-threshing, 126 for seed width, 251 for seed length, and 101 for spikelets per spike - further adding to the argument for polygenic control for these traits.

Brittle rachis and floret shattering were both positively correlated with 10-spike seed weight, and floret shattering was positively correlated with TGW, seed area, width, and length, meaning that genets that shatter more are also more likely to yield more. This is an unfortunate breeding challenge, albeit not a surprising finding, as heavier seeds and spikes exert more force and thus are more likely to shatter. Of the 4 QTL for average shattering, 4 for brittle rachis, and 1 for floret shattering, none were in common with one another. And fortunately, none of these QTL were associated with those controlling seed and spike architecture traits. Brittle rachis and floret shattering have previously been documented as being under differential genetic control (Altendorf et al., 2021a), findings which were upheld in this study. Since average shattering is

the average of the brittle rachis and floret shattering scores, the unique QTL detected for average shattering are most likely QTL that are either associated with floret shattering or brittle rachis or potentially affect both floret shattering and brittle rachis.

Comparisons to other IWG SNPs found in previous studies were difficult, as most relevant studies use previous versions of the IWG genome assembly, and this study used the most recent version, V3.1, thus the chromosome positioning and naming were different for each version. Conversions from V2.1 chromosome names to V3.1 names can be found in Supplemental Table S6. For brittle rachis, the chromosome J03 SNP explained a large portion (16.3%) of phenotypic variation in this population (Table 2.2). An association mapping study by Crain et al. (2022) also identified this SNP, J03_171120774, as important as it explained 1.4% of the phenotypic variation in their population. Moreover, in a nested association mapping study by Altendorf et al. (2021a), they found 4 QTL common in multiple environments with association analysis for brittle rachis across chromosomes 7, 8, 14, and 18 (IWG V2.1), which translates to chromosomes J03, S03, J05, and J06 (IWG V3.1), explaining an average of 1.5% of the phenotypic variation. Altendorf et al.(2021a) also found one QTL in common between their two types of mapping on chromosome 8/S03 that explained up to 23% of phenotypic variation. IWG chromosomes J03 and S03 are part of homeologous group 3, which has demonstrated collinearity to barley chromosome 3H, the known location of 2 brittle rachis genes (*Btr1* and *Btr2*, DeHaan et al., 2020), thus, it is likely the identified markers on these chromosomes could be in close proximity to these orthologous brittle rachis genes in IWG, and would be a good candidate for further study and intentional selection in the breeding program.

Only one locus, on chromosome V05, was found significant for floret shattering in this study accounting for 14.7% of phenotypic variance. Altendorf et al. (2021a) suggested floret

shattering was under polygenic control and found 10 QTL across 8 different chromosomes associated with floret shattering, one of which was on chromosome 15 (V05 in V3.1), which explained a large amount of phenotypic variance (22%). Four loci were identified for average shattering in this study, none of which were found by Crain et al. (2022) for brittle rachis or floret shattering, but 3 were on the same chromosomes, J02 and S01. The locus on J02 was within 2.8Mb of the SNP found in their study.

Favorable allele frequency changes across breeding cycles varied for each trait and SNP. In a previous study comparing UMN IWG cycle 1 and cycle 3 materials, favorable allele frequency increased from cycle 1 to cycle 3 in 71% of individuals, although the increase was not statistically significant (Bajgain et al., 2019a). We expected nearly all loci to increase in favorable allele frequency from cycle 2 to cycle 5 due to selection, but that was not always the case. In the case of shattering, one potential explanation for this unexpected trend could be that brittle rachis and floret shattering are not directly selected in the UMN IWG breeding program, instead, a rating of "average shattering" that considers brittle rachis and floret shattering is used. Of the 4 average shattering loci identified, 3 increased in favorable allele frequency from cycle 2 to cycle 5. This trend is also evident in a previous study of the UMN IWG breeding program, which did not detect significant genetic gain per cycle for brittle rachis shattering, but did see significant improvements for both floret and average shattering across breeding cycles (Stoll et al., 2024, in press). Similarly, for 10-spike seed and spike weight, favorable allele frequencies decreased for the majority of the significant QTL, but favorable allele frequency increased for the majority of spikelet density QTL and the seed width QTL. These findings align with a previous study, which identified a significant decline in 10-spike yield traits with later cycles and a significant increase in seed area, seed width, and spikelet density, despite spikelet density not

being directly selected upon (Stoll et al., 2024, in press). Furthermore, the alignment between the findings in this study and the previous study by Stoll et al. (2024) may suggest the unexpected decline in favorable allele frequency for 10-spike seed and spike weight and brittle rachis should be further investigated. It is possible unfavorable alleles are unintentionally being selected in the genomic selection-based breeding program. And more intentional direct selection on favorable alleles might be justified.

These favorable allele frequency findings were somewhat disappointing but come with multiple caveats. First, there are likely many more QTL for each trait than the few analyzed in this study for allele frequency. This means there could be some inherent sampling error in the 33 identified QTL. Moreover, because there are numerous loci for each trait, making extensive progress for every locus over just four breeding cycles would not be a realistic expectation. This issue is further complicated by IWG's obligate outcrossing nature, which contributes to a high level of individual heterozygosity and heterogeneity in the breeding population. To illustrate this point, each successive IWG breeding cycle is created by allowing selected parents to intercross in a greenhouse, then harvested seeds from the greenhouse intercrossing are planted and used for the next breeding cycle. This system perpetuates a high level of allelic diversity and heterozygosity in the breeding population, makes fixing alleles in the breeding population more challenging than if inbreeding were possible. Furthered by the polygenic nature of most IWG traits, progress can be challenging. Finally, fluctuations in favorable allele frequency across cycles could be due to genetic drift, rather than direct selection, especially since genomic selection rather than a marker-assisted selection approach is used in the breeding program (Zhang et al., 2016; Bajgain et al., 2019a).

2.7 Conclusion

In this study, we completed a GWAS for numerous agronomic and domestication traits in a population of IWG parent plants from four UMN breeding cycles and identified 33 significant QTL. Multiple shattering and brittle rachis QTL were on the same chromosome or close to QTL previously identified in other IWG studies, drawing attention to genomic regions which should be selection targets. The large PVE explained by these QTL is likely a function of the small population size in this study. However, we suggest directly selecting on these identified markers in the IWG breeding program, as incorporation of significant QTL has previously been proven to improve GS models (Bajgain et al., 2019a). Polygenic traits and the obligate outcrossing nature of IWG may hinder breeding progress, but the application of these identified QTL into genomic selection models will help, leading the community one step closer to IWG being an extensively grown perennial grain crop.

2.8 References

Altendorf, K.R., L.R. DeHaan, S.R. Larson, and J.A. Anderson. 2021a. QTL for seed shattering and threshability in intermediate wheatgrass align closely with well-studied orthologs from wheat, barley, and rice. The Plant Genome 14(3): e20145. doi: 10.1002/TPG2.20145.

Altendorf, K.R., S. Larson, L.R. DeHaan, J. Crain, J. Neyhart, et al. 2021b. Nested association mapping reveals the genetic architecture of spike emergence and anthesis timing in intermediate wheatgrass. G3 Genes|Genomes|Genetics. doi: 10.1093/g3journal/jkab025.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3): 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Bajgain, P., and J.A. Anderson. 2021. Multi-Allelic Haplotype-Based Association Analysis Identifies Genomic Regions Controlling Domestication Traits in Intermediate Wheatgrass. Agriculture 2021, Vol. 11, Page 667 11(7): 667. doi: 10.3390/AGRICULTURE11070667.

Bajgain, P., Y. Brandvain, and J.A. Anderson. 2022a. Influence of Pollen Dispersal and Mating Pattern in Domestication of Intermediate Wheatgrass, a Novel Perennial Food Crop. Front. Plant Sci. 13. doi: 10.3389/fpls.2022.871130.

Bajgain, P., C. Li, and J.A. Anderson. 2022b. Genome-wide association mapping and genomic prediction for kernel color traits in intermediate wheatgrass (Thinopyrum intermedium). BMC Plant Biol 22(1): 218. doi: 10.1186/s12870-022-03616-7.

Bajgain, P., X. Zhang, and J.A. Anderson. 2019a. Genome-wide association study of yield component traits in intermediate wheatgrass and implications in genomic selection and breeding. G3: Genes, Genomes, Genetics 9(8): 2429–2439. doi: 10.1534/g3.119.400073.

Bajgain, P., X. Zhang, J.M. Jungers, L.R. DeHaan, B. Heim, et al. 2020. 'MN-Clearwater', the first food-grade intermediate wheatgrass (Kernza perennial grain) cultivar. Journal of Plant Registrations 14(3): 288–297. doi: 10.1002/PLR2.20042.

Bajgain, P., X. Zhang, M.K. Turner, R.D. Curland, B. Heim, et al. 2019b. Characterization of Genetic Resistance to Fusarium Head Blight and Bacterial Leaf Streak in Intermediate Wheatgrass (Thinopyrum intermedium). Agronomy 9(8): 429. doi: 10.3390/agronomy9080429.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67: 1–48. doi: 10.18637/jss.v067.i01.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples. BIOINFORMATICS APPLICATIONS 23: 2633–2635. doi: 10.1093/bioinformatics/btm308.

Broman, K.W., and Ś. Sen. 2009. Fit and exploration of multiple-QTL models. In: Broman, K.W. and Sen, S., editors, A Guide to QTL Mapping with R/qtl. Springer, New York, NY. p. 241–282

Chapman, E.A., H.C. Thomsen, S. Tulloch, P.M.P. Correia, G. Luo, et al. 2022. Perennials as Future Grain Crops: Opportunities and Challenges. Front. Plant Sci. 13. doi: 10.3389/fpls.2022.898769.

Cox, T.S., M. Bender, C. Picone, D.L.V. Tassel, J.B. Holland, et al. 2002. Breeding Perennial Grain Crops. Critical Reviews in Plant Sciences 21(2): 59–91. doi: 10.1080/0735-260291044188.

Cox, T.S., J.D. Glover, D.L. Van Tassel, C.M. Cox, and L.R. DeHaan. 2006. Prospects for Developing Perennial Grain Crops. BioScience 56(8): 649–659. doi: 10.1641/0006-3568(2006)56[649:pfdpgc]2.0.co;2.

Crain, J., P. Bajgain, J. Anderson, X. Zhang, L. DeHaan, et al. 2020. Enhancing Crop Domestication Through Genomic Selection, a Case Study of Intermediate Wheatgrass. Frontiers in Plant Science 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.00319 (accessed 15 May 2023).

Crain, J., S. Larson, K. Dorn, L. DeHaan, and J. Poland. 2022. Genetic architecture and QTL selection response for Kernza perennial grain domestication traits. Theor Appl Genet 135(8): 2769–2784. doi: 10.1007/s00122-022-04148-2.

Crews, T.E., W. Carton, and L. Olsson. 2018. Is the future of agriculture perennial? Imperatives and opportunities to reinvent agriculture by shifting from annual monocultures to perennial polycultures. Global Sustainability 1. doi: 10.1017/sus.2018.11.

Crews, T.E., and D.J. Cattani. 2018. Strategies, Advances, and Challenges in Breeding Perennial Grain Crops. Sustainability 10(7): 2192. doi: 10.3390/su10072192.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, et al. 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science 22(11): 961–975. doi: 10.1016/j.tplants.2017.08.011.

Danecek, P., J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10(2): giab008. doi: 10.1093/gigascience/giab008.

DeHaan, L., M. Christians, J. Crain, and J. Poland. 2018. Development and Evolution of an Intermediate Wheatgrass Domestication Program. Sustainability 2018, Vol. 10, Page 1499 10(5): 1499. doi: 10.3390/SU10051499.

DeHaan, L., S. Larson, R.L. López-Marqués, S. Wenkel, C. Gao, et al. 2020. Roadmap for Accelerated Domestication of an Emerging Perennial Grain Crop. Trends in Plant Science 25(6): 525–537. doi: 10.1016/J.TPLANTS.2020.02.004.

DeHaan, L.R., D.L. Van Tassel, J.A. Anderson, S.R. Asselin, R. Barnes, et al. 2016. A Pipeline Strategy for Grain Crop Domestication. Crop Science 56(3): 917–930. doi: 10.2135/cropsci2015.06.0356.

Dudley, J. w. 2007. From Means to QTL: The Illinois Long-Term Selection Experiment as a Case Study in Quantitative Genetics. Crop Science 47(S3): S-20-S-31. doi: 10.2135/cropsci2007.04.0003IPBS.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L. Orban, editor). PLoS ONE 6(5): e19379. doi: 10.1371/journal.pone.0019379.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54: 357–374. doi: 10.1146/annurev.arplant.54.031902.134907.

Food and Agriculture Organization of the United Nations. (2020, April 10). Global temperature increases by the numbers. FAO. https://www.fao.org/sustainability/news/detail/en/c/1270570/

Hill, W.G., and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33(1): 54–78. doi: 10.1016/0040-5809(88)90004-4.

Jackson, W. 1980. New roots for agriculture. University of Nebraska Press.

Thinopyrum intermedium Genome Sequencing Consortium. Thinopyrum intermedium v3.1 DOE-JGI, https://phytozome-next.jgi.doe.gov/info/Tintermedium_v3_1

Krug, A.S., E. B. M. Drummond, D.L. Van Tassel, and E.J. Warschefsky. 2023. The next era of crop domestication starts now. Proceedings of the National Academy of Sciences 120(14): e2205769120. doi: 10.1073/pnas.2205769120.

Larson, S., L. DeHaan, J. Poland, X. Zhang, K. Dorn, et al. 2019. Genome mapping of quantitative trait loci (QTL) controlling domestication traits of intermediate wheatgrass (Thinopyrum intermedium). Theoretical and Applied Genetics 132(8): 2325–2351. doi: 10.1007/s00122-019-03357-6.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14): 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al. 2009. The Sequence Alignment/Map format and SAMtools. BIOINFORMATICS APPLICATIONS NOTE 25(16): 2078–2079. doi: 10.1093/bioinformatics/btp352.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, et al. 2012. Genetics and population analysis GAPIT: genome association and prediction integrated tool. 28(18): 2397–2399. doi: 10.1093/bioinformatics/bts444.

Liu, X., M. Huang, B. Fan, E.S. Buckler, and Z. Zhang. 2016. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genetics 12(2): e1005767. doi: 10.1371/journal.pgen.1005767.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819–1829. doi: 10.1093/genetics/157.4.1819.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.Y. Zhong, et al. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genetics 5(11): 2383–2390. doi: 10.1534/g3.115.021667.

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLOS ONE 7(2): e32253. doi: 10.1371/journal.pone.0032253.

R Core Team. 2021. R: A language and environment for statistical computing. https://www.R-project.org/.

Stoll, H., Bajgain, P., & Anderson, J. (in press). Assessing genetic gain in an intermediate wheatgrass improvement program: A retrospective analysis. *Crop Science*.

Van Tassel, D.L., L.R. DeHaan, and T.S. Cox. 2010. Missing domesticated plant forms: can artificial selection fill the gap? Evol Appl 3(5–6): 434–452. doi: 10.1111/j.1752-4571.2010.00132.x.

Vogel, K.P., K. Arumuganathan, and K.B. Jensen. 1999. Nuclear DNA Content of Perennial Grasses of the Triticeae. Crop Science 39(3): cropsci1999.0011183X003900020009x. doi: 10.2135/cropsci1999.0011183X003900020009x.

Wagoner, P. 1990. Perennial grain new use for intermediate wheatgrass. Journal of Soil and Water Conservation 45(1): 81–82.

Wagoner, P., and J.R. Schaeffer. 1990. Perennial grain development: Past efforts and potential for the future. Critical Reviews in Plant Sciences 9(5): 381–408. doi: 10.1080/07352689009382298.

Weir, B.S. 2008. Linkage Disequilibrium and Association Mapping. Annual Review of Genomics and Human Genetics 9(1): 129–142. doi: 10.1146/annurev.genom.9.081307.164347.

Zhang, X., S.R. Larson, L. Gao, S.L. Teh, L.R. DeHaan, et al. 2017. Uncovering the Genetic Architecture of Seed Weight and Size in Intermediate Wheatgrass through Linkage and Association Mapping. The Plant Genome 10(3): plantgenome2017.03.0022. doi: 10.3835/plantgenome2017.03.0022.

Zhang, X., A. Sallam, L. Gao, T. Kantarski, J. Poland, et al. 2016. Establishment and Optimization of Genomic Selection to Accelerate the Domestication and Improvement of Intermediate Wheatgrass. The Plant Genome 9(1): plantgenome2015.07.0059. doi: 10.3835/plantgenome2015.07.0059.

3. Teaching with Kernza®: Curriculum Development for Biological Concepts and Plant Breeding Using Applied Agricultural Research

3.1 Abstract

The demand for professionals trained in science, technology, engineering, and mathematics (STEM), and agriculture has steadily increased for multiple decades and is expected to continue to increase. Simultaneously, discrepancies in quantitative skills development and exposure to STEM and agriculture career paths at the high school and undergraduate levels have become apparent. This work addresses the demand for curricula integrating biological sciences and agriculture while prioritizing quantitative skills and active learning methods. The developed Kernza® in Context Curriculum exposes students to the social, economic, and ecological implications of Kernza® as a new, sustainable perennial grain crop. In this work specifically, a Plant Breeding and Genetics Module consisting of 3 individual lessons and two additional Data Nugget activities was developed as a component of a broad curriculum. The curriculum was developed using a backward design and curriculum mapping approach, then iteratively beta-tested and revised. From here, we make recommendations based on our own experiences and additional literature review to propose a 3-part approach to curriculum development and evaluation: 1) curriculum mapping, 2), collecting feedback, and 3) additional research.

3.2 Introduction

Research findings in the academic world are typically shared via journal articles, at conferences, and in the classroom. These modes of sharing research findings may be considered effective for sharing knowledge amongst researchers, but they largely exclude the general public. Moreover, research dissemination in the classroom (e.g. in an undergraduate classroom) is often

difficult due to faculty with research appointments allowed limited time to develop a comprehensive curriculum, and educators at institutions without research appointments may have limited access to relevant curricula in more specific fields. Additionally, journal articles, conferences, and university classrooms require large fees to access, further limiting their viability as a means for sharing scientific findings outside of academia. Other science-communication avenues exist, such as extension publications in the case of agricultural research, podcasts, social media, and other communications-type journals (Clifton-Ross et al., 2019; Figueroa, 2022). However, the existence of relevant, openly accessible curricula informed by research findings is largely lacking.

STEM, which stands for "Science, Technology, Engineering, and Mathematics" began as a conceptualization of the National Science Foundation (NSF) in the early 1990s, and broadly refers to mathematics, engineering, computer sciences, natural sciences, and social/behavioral sciences. Since the 1990s, STEM concepts have been integrated by numerous local, state, and national educational programs, but the understanding of the implications and what constitutes "STEM" varies (Breiner et al., 2012). Many organizations recognize the importance of STEM education for future careers, as occupations in STEM are expected to increase by over 10% from 2022 to 2032, in comparison to an expected increase of 2.3% for non-STEM careers in the same timeframe (Employment in STEM Occupations, 2024). Additionally, careers in agriculture and food science are expected to increase by 6% in the same timeframe (Occupational Outlook Handbook: Agricultural and Food Scientists, 2024).

Both policy (NSF and legislative organizations) and educational organizations largely view STEM as a set of disciplinary coursework in science, technology, engineering, and mathematics, without substantial integration of those fields. Consequently, many today criticize

the lack of integration of those four disciplines in the modern U.S. educational system, and how vital their integration is for students to confidently apply STEM concepts to solve real-world problems (Breiner et al., 2012; Labov et al., 2010). Specifically for the biological sciences, Bialek & Botstein argue that the biological sciences are largely excluded from the physical sciences and engineering, which better incorporate quantitative and mathematical skills development. Even at the post-secondary level, students in STEM are often divided into 2 different "tracks"; the first being those who wish to pursue a career in the biological sciences or medical fields, and the second being those interested in mathematics, chemistry, physics, and engineering. At the post-secondary level, these two tracks of students are usually taught the basic STEM classes separately, and the curricula for the students in the biological sciences-related track usually lacks major integration and quality of quantitative skills (Bialek & Botstein, 2004). Moreover, mathematics and biological sciences courses are usually taught completely separately, leaving students with two separate pools of knowledge, and limited opportunities to integrate and synthesize those disciplines to problem-solve (Gross, 2000, p. 200).

Feser et al. describes two case studies in undergraduate biology courses aimed at integrating quantitative skills into the biology classroom (Feser et al., 2013). In the first case study, a set of over 40 short "MathBench Biology Modules" were created and utilized to teach concepts such as graphing, statistics, metric conversions, and more. These were easily integrated into existing biology curriculum and could also be accessed by students on an as-needed basis. Educators reported improvements in quantitative skills, assessed by pre- and post-tests, improved student engagement, and understanding of the integration of math and biology (Thompson et al., 2010). Moreover, in a second case study, the researchers rewrote a textbook to minimize memorization and prioritize the integration of data interpretation and math in biology. Despite

reducing biological sciences content, students retained a similar amount of content with the old and redesigned textbooks. However, students retained more knowledge long-term with the new text. Moreover, students taught with the new textbook were better able to interpret data (Barsoum et al., 2013). Other research has demonstrated how intentional integration of module-based activities in Mendelian genetics, mathematical modeling, and population genetics in a biology class improves quantitative numeracy skills, data interpretation, and ability to make biological inferences (Hoffman et al., 2017). Thus, a significant body of research demonstrates that module-based and other activities aimed at practicing quantitative skills in undergraduate biology courses lead to improved quantitative reasoning student outcomes.

In direct response to this lack of integrative quantitative reasoning curricula in the life sciences, "Data Nuggets" was created. Data Nuggets (https://datanuggets.org) is a free platform of curricula designed to bring real research and datasets into the K-16 classroom to enhance students' quantitative and inquiry skills (Schultheis & Kjelvik, 2015). Data Nuggets activities are created through an iterative review process involving both teachers and scientists. Research has demonstrated that incorporating Data Nuggets activities into existing classroom curricula leads to students who are more engaged in science, and able to create logical scientific explanations for problems (Schultheis et al., 2023). Moreover, students who engage with Data Nuggets in the classroom express more interest in STEM career paths and exhibit more independence with data manipulation and analysis (Schultheis et al., 2023).

Finally, effective curricula that achieve desired learning outcomes can be developed through strategic design. Backward design is an effective approach to curriculum development, as it prioritizes learning objectives over the content to be covered, thus improving students' understanding of topics. Backward design begins by first outlining the desired results for a lesson

(i.e. intended learning outcomes), then determining the assessment techniques, and finally planning learning activities and lessons. The learning outcomes must be specific, measurable, realistic, and student-centered, and the assessments and lesson plan must be aligned with the originally outlined learning outcomes (Wiggins & McTighe, 2005). In addition, effective curricula should design in-class activities that promote active learning. Active learning is defined as "instructional activities involving students in doing things and thinking about what they are doing" (Bonwell & Eison, 1991). Active learning integration has been proven to promote more equitable, student-centered classrooms that achieve desired learning outcomes (Michael, 2006). Moreover, the incorporation of 3-dimensional learning (3DL), an alternative approach to STEM course transformation, encourages students to apply their learned knowledge to explain, model, and predict phenomena. 3DL improves student engagement specifically with the core ideas of the discipline (Cooper et al., 2024).

It is evident that sufficient curricula which expose students in the agricultural and biological sciences to quantitative skills development are not available for educators at multiple levels. Moreover, effective techniques to assess the effectiveness of any curricula developed by biological and agricultural researchers to meet these educational needs has not been investigated. Thus, the objectives of this work were to 1) demonstrate two methods for agricultural and biological researchers to disseminate their research to improve quantitative educational outcomes for K-16 students and increase student exposure to careers in agriculture and STEM, and 2) provide suggestions to educators on the evaluation and continual modification of developed curricula based on a literature review and learned experiences developing curriculum for this purpose.

3.3 Curriculum Development Procedure

3.3.1 Kernza® in Context - Breeding and Genetics Module

The Kernza® in Context Curriculum is an open-access curriculum developed as part of the Kernza® Coordinated Agricultural Project (KernzaCAP) grant, funded in 2020. The main learning objectives of Kernza® in Context are to help students 1) understand Kernza® as a perennial grain crop, 2) explore the social, economic, and ecological implications of establishing a new perennial grain crop, and 3) understand agriculture in an ecological context. This curriculum prioritizes hands-on activities that can be taught in one or a few class periods. The curriculum is based on modules and lessons that align with national and state curricular standards for high schools, allowing high school educators to pick out lessons that meet their classroom needs and ensure they are completing all required teachings for their subject area. Lessons within the Kernza® in Context Curriculum were developed by scientists from around the country - many were from The Land Institute (Salina, KS) or directly involved with the KernzaCAP grant. This work will showcase a Kernza® Breeding and Genetics Module I developed as a part of this holistic curriculum.

The Kernza® Breeding and Genetics Module is separated into four independent lessons, the last lesson being a Data Nugget, which is described in further detail in the next section (Figure 3.1). Each lesson was designed using a "Backward Design" approach to prioritize desired student learning outcomes, rather than specific topics, and build activities around those desired learning outcomes (Wiggins & McTighe, 2005). First, learning outcomes for the module as a whole and each lesson were determined. Then, identified learning objectives were aligned with NGSS standards (https://www.nextgenscience.org/) to ensure materials were relevant for

educators in the K-12 setting. Moreover, “Essential Questions” for each lesson were determined to ensure alignment with desired learning outcomes. Finally, hands-on in-class activities were designed to target the Essential Questions and learning objectives for each lesson. Each lesson includes guided instructions and a slide deck for the educator to easily implement the activity in their classroom, a list of key vocabulary words (to reference if needed), an assessment tool (typically a short quiz), a rubric to grade the assessment, and other worksheets or printable activities, depending on the lesson.

3.3.2 Data Nugget - “A plant breeder’s quest to improve perennial grain.”

The Data Nugget was created using the “Experimental study” template from the Data Nuggets website (https://datanuggets.org/making-your-own-nugget/). Data Nuggets may be targeted for the K-16 audience; this Data Nugget targets high school and undergraduate-level learners. The experimental template includes a research background description, scientific question, and a hypothesis statement (which students must identify in-text). For this developed Data Nugget, scientific data in the format of tables were included for students to use to answer the scientific question using graphs and other fillable tables. There are 3 types of graphing activities educators can choose to use in their classroom (Figure 3.2), all of which should be provided in the Experimental study submission. In Graph Type A, a graph is provided and students must interpret the results to answer the scientific question. In Graph Type B, axis labels and scale are provided, but the student graphs the data. And in Type C, the student must graph the data without any context on axis labels, scales, or an existing graph. After graphing the data and making any necessary calculations, the students answer a series of questions to interpret the

data, determine if the data support the researcher's hypothesis, explain why or why not, and propose some new questions to be investigated.

In this particular Data Nugget, "A plant breeder's quest to improve perennial grain", the goal of the activity is to emulate the process a scientist would follow when performing genomic selection in a breeding program (Jannink et al., 2010). Thus, datasets include training data, test data, and validation data to address two hypotheses: 1) "If a genotype leads to a phenotype, then she can use the combined phenotype and genotype information from the first four cycles to predict the average phenotype of each genotype" and 2) "If she can use Cycle 1-4 predicted phenotypic effects to predict the 800 un-phenotyped Cycle 5 plants, then the difference between the Cycle 1-4 phenotypic effects and the Cycle 5 validation dataset (the 200 phenotyped plants) should not differ significantly." These two hypotheses were separated into a main activity and a supplemental activity; Data Nugget Activity worksheets, datasets, and graphs can be found in Chapter 3 Supplemental Information.

3.3.3 - Curriculum Evaluation

The Kernza® in Context Curriculum and Data Nugget were evaluated using multiple qualitative methods to assess the effectiveness of the modules and activities with the desired learning outcomes. First, the alignment of the lessons within the Breeding and Genetics module of the Kernza® in Context Curriculum were mapped using a curriculum mapping approach (Allen, 2004). In short, this mapping approach was used to identify any potential gaps in expected student learning outcomes and what was taught in each lesson. For each lesson, it was determined whether the lesson introduced, reinforced, mastered, or assessed each learning outcome. Introduction of a learning outcome involves simply introducing the student to the

concept, potentially with an activity or other content-delivery method. Next, reinforcement gives students the opportunity to further practice the desired learning outcome, potentially through formative assessment such as a low-stakes in-class activity. Mastery may also involve formative assessment, with the goal of the student understanding the topic at the senior or exit level. Finally, students may be assessed formatively or summatively (i.e. with an exam) to evaluate success of teaching the intended learning outcomes.

The Kernza® in Context Curriculum was reviewed by peer scientists for general edits and comments concerning the content and formatting in 2022. The Kernza® in Context Curriculum was beta-tested by a team of high school and college educators in 2022 and 2023. These educators reviewed the materials to make suggestions for improvements to activities and to improve the utility and accessibility of each lesson. Some of the lessons were directly taught in educators' classrooms, where educators then filled out a post-lesson survey to make suggestions for "things that could be improved" and "things that went well" with each lesson. After this first round of beta testing, adjustments to each lesson were made. In addition to in-class beta-testing by educators in 2022 and 2023, the Data Nugget was also beta-tested by fellow colleagues and scientists at the 2023 KernzaCAP meeting. At this meeting, around 25 people completed the lesson activities and shared feedback in both written and verbal form. Lastly, the Data Nugget and one Kernza® in Context Lesson: "Tools in a Plant Breeder's Toolbox: Genotyping and Phenotyping Technology" were beta tested with community college and high school educators at the Wisconsin Association of Agricultural Educators meeting in Oshkosh, Wisconsin in 2023. At this meeting, additional verbal feedback was given about aspects of the lessons that the educators foresaw as working well with their students, and suggestions for the best method of publishing the curricula for educator access was shared.

Extensive peer review and beta-testing, rather than explicitly quantitative methods for curriculum assessment were used in this approach. Therefore, a brief literature review will be performed to make suggestions for future quantitative methods that could be used to assess the effectiveness of a developed curriculum similar to the modules presented in this work. From this, we propose a framework for curriculum evaluation for similar curriculum development efforts in the future, including curriculum mapping, collecting feedback, and research.

3.4 Curriculum Development Outcomes

3.4.1 Kernza® in Context - Breeding and Genetics Module

The Kernza® in Context curriculum resulted in 10 modules with a total of 26 lessons nested within different modules. Module topics outside of plant breeding and genetics included topics such as climate change, agricultural supply chains, research and data, and multifunctional cropping systems. Lessons touched on baking with Kernza®, plant physiology, the history of grain domestication, livestock integration, soil carbon, ecosystem services, multiple writing activities about perennial agricultural systems, and more. The published curriculum can be found and openly accessed at https://drive.google.com/drive/folders/1auLD4bnkFMDPEaI607goO-eIqTiVS_cH.

The Plant Breeding and Genetics module comprised 4 lessons, the last of which is the Data Nugget described in further detail later. Concurrent with the backward design used to design these modules, teachers are provided with a document outlining the Main Idea, Driving Questions, Lesson Outcomes, and Next Generation Science Standards. A snapshot example of the information provided for each lesson is provided in Figure 3.3. Additionally, the lesson contents and a teaching outline are provided (Figure 3.4). The complete module including slide

decks, activities, full teaching outlines, rubrics, vocabulary sheets, and assessments can be found in the Chapter 3 Supplemental Information.

3.4.2 Data Nugget - "A plant breeder's quest to improve perennial grain."

This work resulted in a two-part Data Nugget activity for high school and undergraduate-level learners. Lesson facilitators can adjust the level of difficulty by using different graph types (Figure 3.2). The hypothesis addressed by the main activity was "If a genotype leads to a phenotype, then she [Hannah] can use the combined phenotype and genotype information from the first four cycles to predict the average phenotype of each genotype." To determine if this statement is true, students were given "training data" that provided seed area and single nucleotide polymorphism (SNP) data for several plants; students had to combine SNP data to identify the genotype of each plant (Table 1a, Chapter 3 Supplemental Information). Then, students had to calculate the average seed area for each genotype (Table 1b, Chapter 3 Supplemental Information). Finally, students were to identify the independent and dependent variables for graphing, which were genotype and average seed area (phenotypic means), respectively (Figure 3.5). After graphing, students were asked a series of questions about what evidence supports their claim and why, whether or not the data support the hypothesis, and what the next steps or additional questions they might ask as scientists after completing this case study. One potential additional question we hope students will ask (and teachers are prompted to propose if they do not) is how a plant breeder could apply these concepts in a real breeding program, which leads students into the second supplemental activity.

The follow-up supplemental activity, "Supplemental Activity: A plant breeder's quest to improve perennial grain" builds on the first activity connecting genotype and phenotype to apply

the predicted phenotype values calculated in the first activity to make predictions about the next breeding cycle, thus emulating how genomic selection works in a real breeding program. The hypothesis addressed by this supplemental activity was "If she [Hannah] can use Cycle 1-4 predicted phenotypic effects to predict the 800 un-phenotyped Cycle 5 plants, then the difference between the Cycle 1-4 phenotypic effects and the Cycle 5 validation dataset (the 200 phenotyped plants) should not differ significantly." To answer this scientific question, they were provided with a schematic of how a breeding program works in the context of the data they were given (Figure 3.6). Then, students were given test data calculated from Cycles 1-4 in the previous activity and validation data from Cycle 5 (Table 2a and 2b, Chapter 3 Supplemental Information). With these data, students were to calculate the difference between the data trained from Cycles 1-4 to the data from Cycle 5 to validate their hypothesis (Table 2c, Chapter 3 Supplemental Information). In addition, they were to graph the Cycle 1-4 test data and Cycle 5 validation data to visualize any obvious differences between the two that would suggest a rejection of the hypothesis (Figure 3.7). The complete activity including datasets and teacher notes can be found in Chapter 3 Supplemental Information.

3.3.3 - Curriculum Evaluation

We propose a three-part curriculum evaluation scheme (Figure 3.8). First, lessons and learning objectives should be mapped to ensure a comprehensive curriculum that accomplishes all desired learning outcomes. Next, feedback from teachers, students, peer researchers, and more should be collected and used to iteratively modify the curriculum. Finally, additional research studies that quantitatively evaluate learning outcomes from the curriculum can be applied.

Prior to any beta-testing, the curriculum was mapped using an approach from Allen (2004) to ensure alignment of learning objectives (Table 3.1, Figure 3.8). Of the selected learning outcomes identified in Table 3.1, lessons within the module build on each learning objective by introducing, reinforcing, mastering, and assessing students on each topic. After the completion of the 4-lesson module, students have mastered multiple topics such as understanding that living things have DNA and DNA leads to protein, calculating summary statistics, and testing a hypothesis (Table 3.1). Moreover, educators have the opportunity to formatively or summatively assess students to ensure they are achieving desired learning outcomes in each lesson through in-class activities (formative assessment) and learning quizzes (summative assessment). While lessons were designed to taught in sequence, teachers are provided with prerequisite information so they can choose to teach just one of the lessons, if they have already covered the intended learning objectives using other curriculum or teaching techniques.

Next, feedback from fellow researchers, and high school and college educators was collected (Figure 3.8). Concerning the qualitative evaluation of the curricula with beta-testing, multiple adjustments were iteratively made to both the Kernza® in Context Breeding and Genetics Module and the Data Nugget after beta-testing. After the first phase of beta-testing of the Kernza® in Context Module by a few high school educators, some changes were made to lesson content and activities. For example, they indicated the “Genes are Life’s Instruction Manual” activity needed to be modified to prevent students from easily cheating and missing the point of the activity (which was to demonstrate DNA transcription and RNA translation). In addition, verbal and written feedback from two Kernza® scientific meetings and the Wisconsin Association of Agriculture Educators Conference were incorporated to improve the robustness and clarity of materials and activities. Speaking to educators directly proved to be one of the

most insightful ways of beta-testing and gathering feedback. Educators shared personal opinions on what would make them more willing to use the developed curriculum in their classroom. Multiple educators emphasized the importance of having engaging, hands-on activities, minimal preparation needed for activity facilitation, shorter lessons that could stand alone to teach one or two key concepts and indicated that sharing the curriculum via Google Drive was the best platform for public school teachers to access the curriculum. For example, in Wisconsin, the Wisconsin Department of Public Instruction has a shared Google Drive with all available curricula (and other states have similar organizational platforms), thus making Google Drive an easy way to share developed curricula with high-school educators. Thus, the Kernza® in Context Curriculum is shared openly online with Google Drive.

Finally, although additional research with extensive quantitative methods was not performed, we propose a few suggestions for a more systematic evaluation of desired learning outcomes based on current literature (Figure 3.8). By and large, assessment is the best method to quantitatively evaluate student learning and the effectiveness of the curriculum. Various formative or summative assessment techniques can be applied and analyzed. Concerning summative assessment, pre-tests and post-tests given after the curriculum is taught can illustrate which learning outcomes are being effectively taught, and illuminate others that potentially need modification (Feser et al., 2013; Hoffman et al., 2017). Exam scores may be evaluated for each desired learning outcome to identify areas of weakness within the curriculum. On a similar note, assessment via pre- and post-testing could be used to compare two different groups of students - those taught with the newly developed curriculum, and those taught with an older version, or other unrelated curriculum. Moreover, assessment beyond summative assessment via tests and quizzes may be useful. Formative assessments that range in knowledge expression, such as

lower-stakes activities, short quizzes, or concept maps can illuminate learning discrepancies (McKeachie, 2014). Using the results from lower-stakes activities may indicate which activities are more effective than others, while better representing the voices of students who may not perform well on exams. For example, for the activity in the "Genes are Life's Instruction Module" lesson where students group together to perform DNA transcription and RNA translation to a protein using printed materials, students and the teacher could be surveyed post-activity to rate their understanding of transcription and translation (e.g. "On a scale of 1 [I do not understand at all] to 5 [I understand completely], please identify your understanding of DNA leading to protein"). Finally, there are ample considerations for performing additional research to evaluate curriculum effectiveness. Practitioners must plan ahead. For studies involving human subjects, appropriate Institutional Review Board (IRB) approval must be obtained. Moreover, appropriate experimental designs with sufficient statistical power to detect differences between student groups must be considered.

3.5 Concluding Remarks

In conclusion, the developed Kernza® in Context Curriculum and Data Nugget address multiple important educational needs. Despite the demand for careers in STEM and agriculture expected to increase, fewer individuals in the United States grow up directly exposed to agriculture. Thus, student exposure to careers in STEM and agriculture through school curricula is one method to address this foreseeable issue. Moreover, curriculum development is another way to authentically disseminate research findings in agriculture and biology, while exposing students to potential career opportunities in those fields. To this end, the integration of authentic research into an openly accessible curriculum may also improve efforts to improve diversity,

equity, and inclusion in the STEM and agricultural fields. Moreover, a substantial amount of research shows that incorporating module-based exercises and similar activities into undergraduate biology courses enhances students' quantitative reasoning skills (Barsoum et al., 2013; Feser et al., 2013; Hoffman et al., 2017; Thompson et al., 2010). Despite these findings, STEM concepts are thought to be poorly integrated into biological curricula in meaningful ways. Thus, this developed curriculum may help mitigate these educational discrepancies by improving student engagement in STEM and agriculture as potential career paths and improving quantitative skills and biological concept retention. Finally, our proposed framework for developing and evaluating curriculum may be used as a guideline for other researchers and educators to incorporate real research and quantitative skills development activities into useful curricula.

3.6 Tables and Figures

Figure 3.1: Kernza® Breeding and Genetics Module Outline

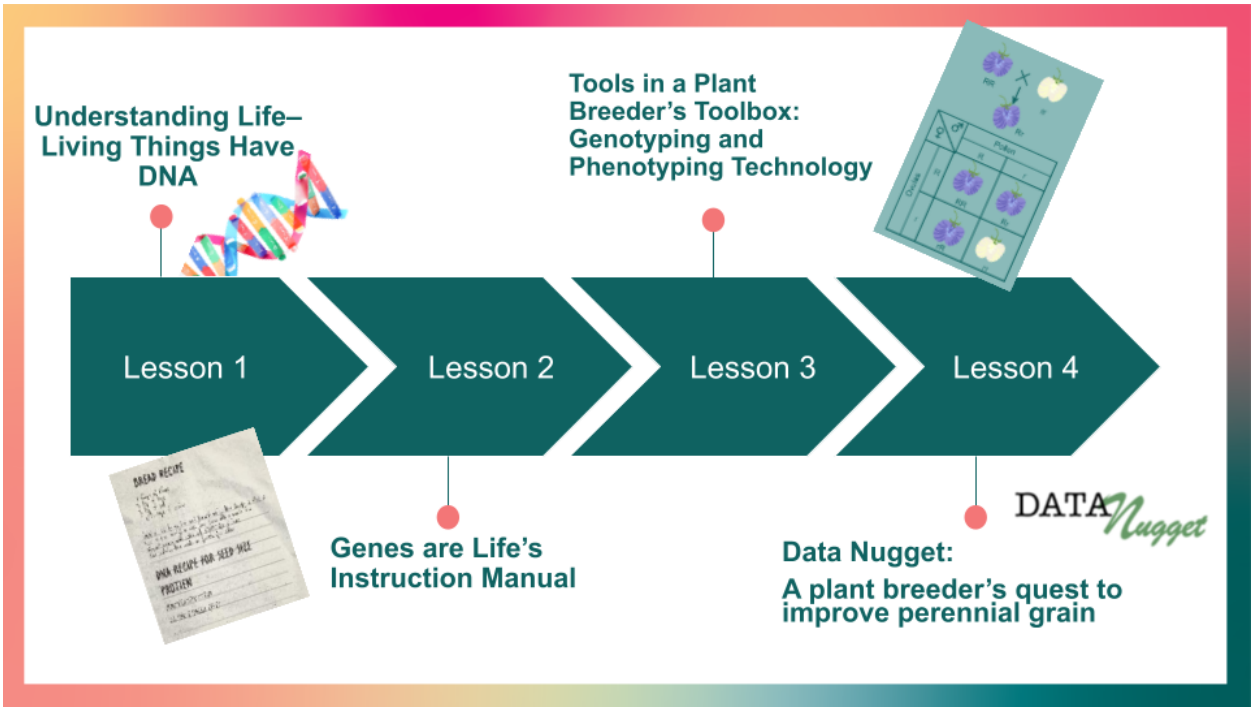
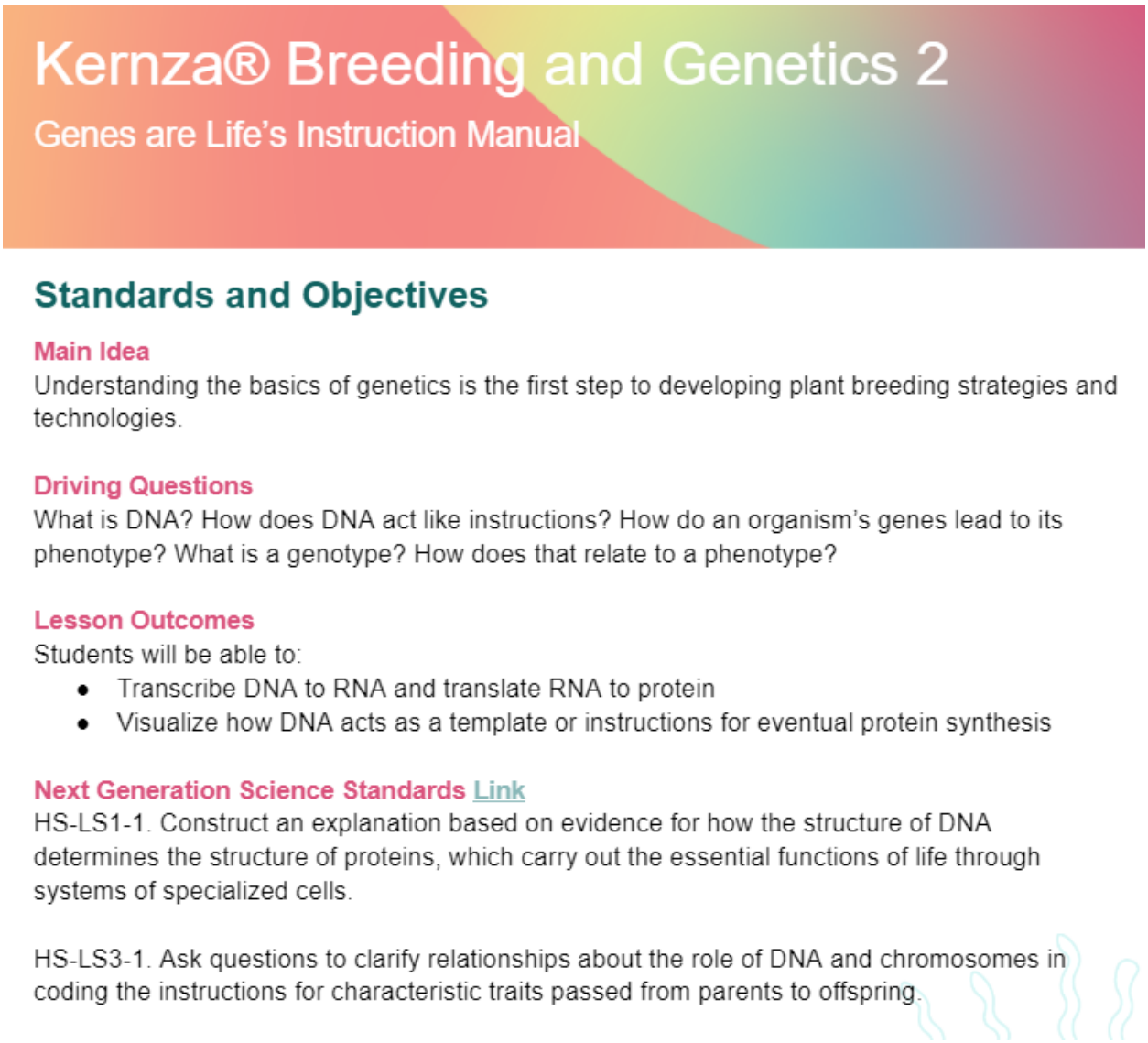


Figure 3.2: Data Nugget Graph Types

Type A graph provided	• Data: displayed on graph • Axis labels and scale provided
Type B	• Data: student graphs data • Axis labels and scale provided
Type C student creates graph	• Data: student graphs data • Axis labels and scale not provided

Figure 3.3: Standards and Objectives provided for each lesson in the Kernza® in Context Curriculum. Backward design is used to first establish the main idea, driving questions, and learning objectives, then these are matched with NGS Standards.



Kernza® Breeding and Genetics 2

Genes are Life's Instruction Manual

Standards and Objectives

Main Idea

Understanding the basics of genetics is the first step to developing plant breeding strategies and technologies.

Driving Questions

What is DNA? How does DNA act like instructions? How do an organism's genes lead to its phenotype? What is a genotype? How does that relate to a phenotype?

Lesson Outcomes

Students will be able to:

- Transcribe DNA to RNA and translate RNA to protein
- Visualize how DNA acts as a template or instructions for eventual protein synthesis

Next Generation Science Standards Link

HS-LS1-1. Construct an explanation based on evidence for how the structure of DNA determines the structure of proteins, which carry out the essential functions of life through systems of specialized cells.

HS-LS3-1. Ask questions to clarify relationships about the role of DNA and chromosomes in coding the instructions for characteristic traits passed from parents to offspring.

Figure 3.4: Teaching Outline provided for each lesson in the Kernza® in Context Curriculum.

Teaching Genes are Life's Instruction Manual

Lesson Components		
Subject: Biology	**Grade Level:** High School	**Lesson Time:** 50-60 min
Supplies Needed: • paper • writing utensils • printed worksheets • scissors	**Activity type:** • Presentation, • Group Discussion • Group activity	**Links to Resources Used:** • N/A
Assessment type: • Pre/Post Assessment	**What Teachers Should Know:** Understand the central dogma of biology: "DNA → RNA → Protein" to help students understand how genotype leads to phenotype.	**What Students Should Know:** Students should have a basic understanding of biology and have completed Lesson 1 or already have a general understanding of life & DNA.

Lesson Outline	
Anticipatory Set/Opening: ***Learning interest "hook" Opening prompts***	*5 minutes: Use the first two slides as a review, asking students what are the characteristics of living things.*
Methods/Procedures:	**Presentation and Discussion (25 Min)** Follow along with the instructions in the presentation. **Build a Protein Activity (15 Min)** Print and cut out the included assignment sheet and follow the instructions to lead the activity.
Closure and Follow-Up:	*Use the included assessment questions.*

Figure 3.5: Graph of Average Seed Area (mm) by Genotype for Data Nuggets Main Activity.

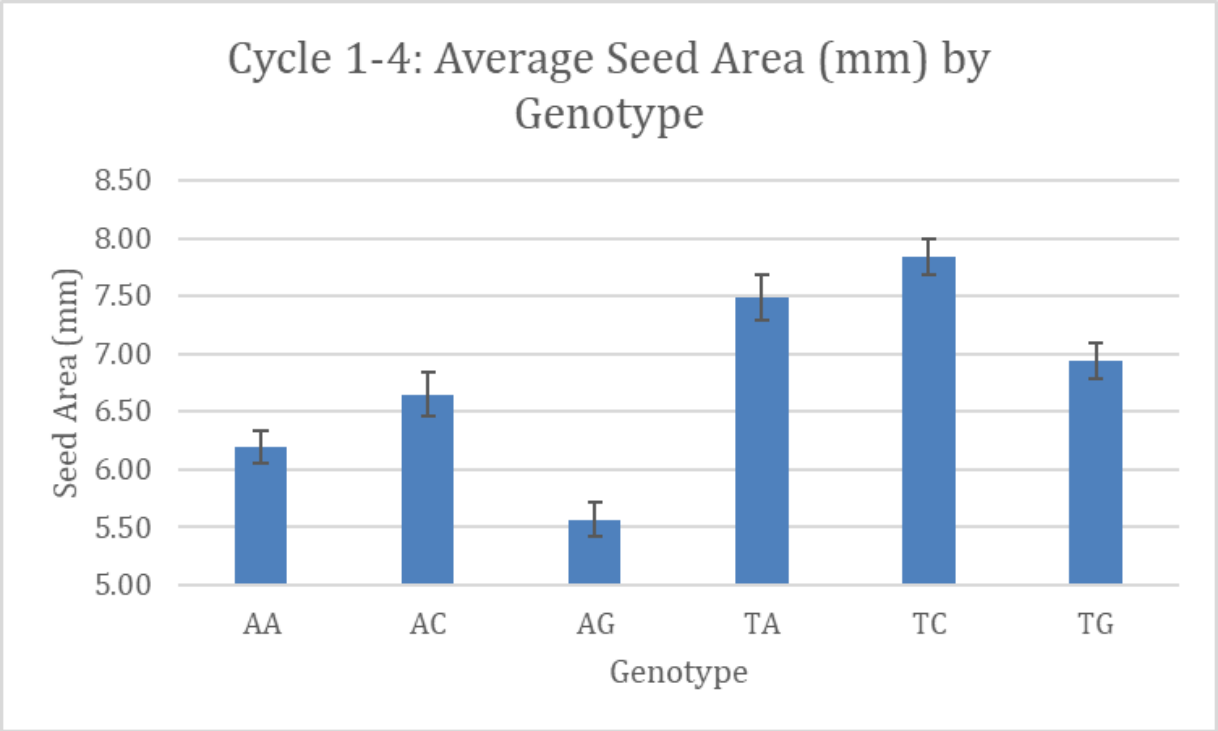


Figure 3.6: Schematic of how a plant breeder may use both phenotypic and genotypic data in a Kernza® breeding program to make selections (i.e. Genomic Selection) for Data Nuggets Supplemental Activity.

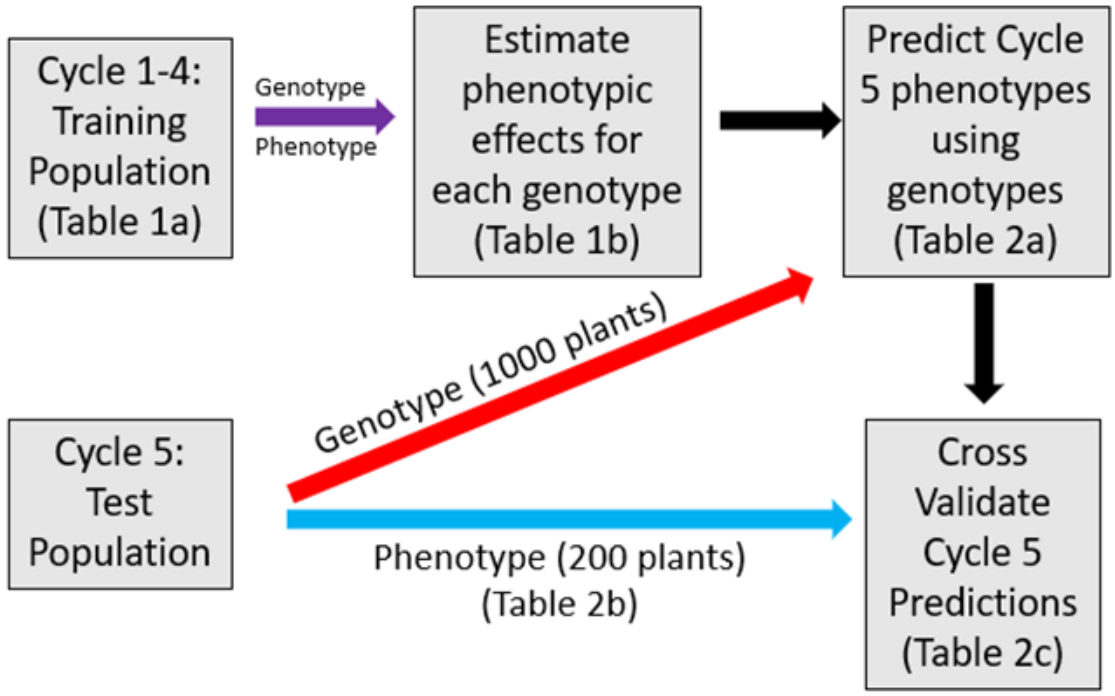


Figure 3.7: Graph of Average Seed Area (mm) by Genotype and Cycle Data Source for Data Nuggets Supplemental Activity.

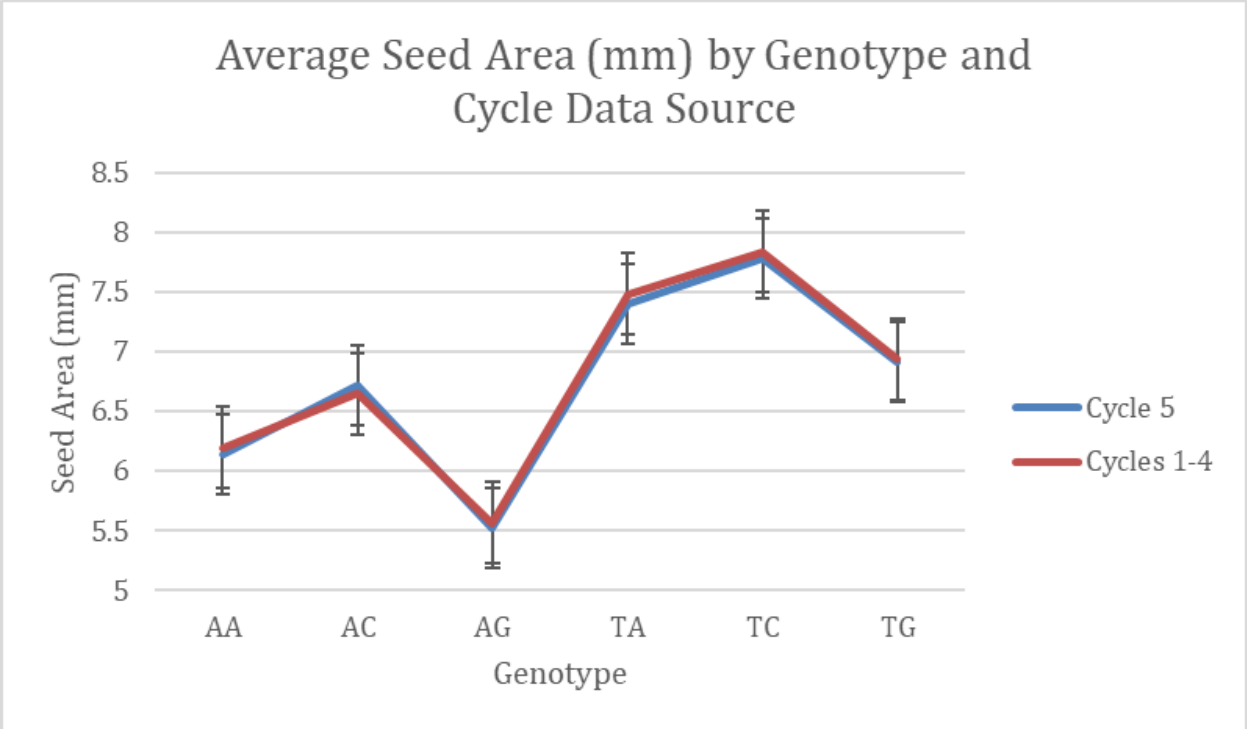


Figure 3.8: Proposed three-part curriculum evaluation scheme: Curriculum Mapping, Collecting Feedback, and Additional Research.

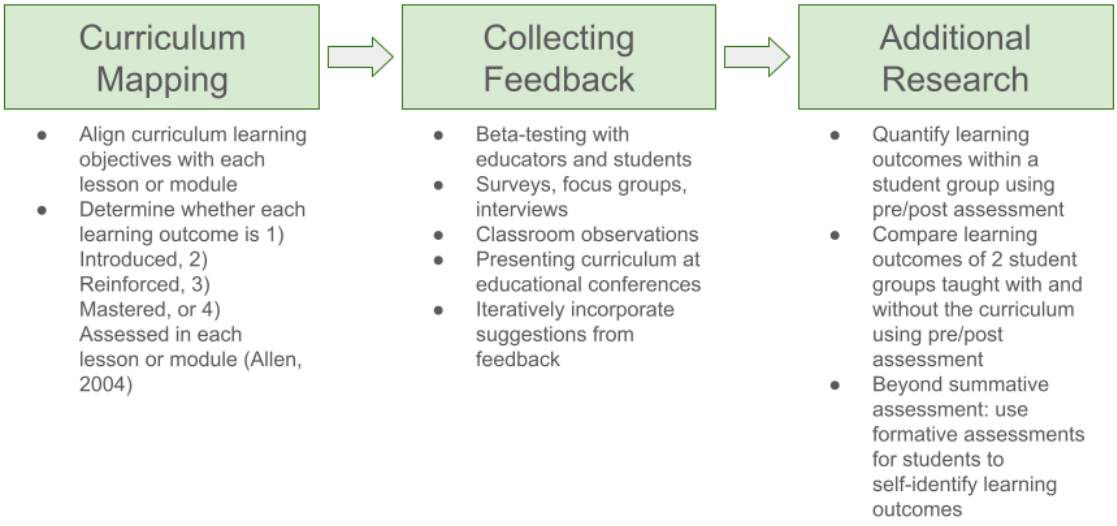


Table 3.1: Example Curriculum Map, modified from Allen, 2004, and the University of Hawaii Manoa, 2013. Categories involve introducing, reinforcing, mastering, and assessing learning outcomes for each lesson.

Lessons	**Learning Outcomes**				
	Living things have DNA	**DNA to Protein**	**Breeding Technologies**	**Calculate Summary Statistics**	**Develop and test a hypothesis**
Understanding Life: Living Things have DNA	Introduce				Introduce
Genes are Life's Instruction Manual	Reinforce	Introduce		Introduce	
Tools in a Plant Breeder's Toolbox: Genotyping and Phenotyping Technology	Reinforce, Master		Introduce	Reinforce	Reinforce

Data Nugget: A plant breeder's quest to improve perennial grain	Assess	Reinforce, Master	Reinforce	Master, Assess	Master, Assess

3.7 References

Allen, M. J. (2004). *Assessing Academic Programs in Higher Education*. John Wiley & Sons.

Barsoum, M. J., Sellers, P. J., Campbell, A. M., Heyer, L. J., & Paradise, C. J. (2013). Implementing Recommendations for Introductory Biology by Writing a New Textbook. *CBE—Life Sciences Education*, *12*(1), 106–116. https://doi.org/10.1187/cbe.12-06-0086

Bialek, W., & Botstein, D. (2004). Introductory Science and Mathematics Education for 21st-Century Biologists. *Science*, *303*(5659), 788–790. https://doi.org/10.1126/science.1095480

Bonwell, C. C., & Eison, J. A. (1991). *Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports*. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183 ($17. https://eric.ed.gov/?id=ED336049

Breiner, J. M., Harkness, S. S., Johnson, C. C., & Koehler, C. M. (2012). What Is STEM? A Discussion About Conceptions of STEM in Education and Partnerships. *School Science and Mathematics*, *112*(1), 3–11. https://doi.org/10.1111/j.1949-8594.2011.00109.x

Clifton-Ross, J., Dale, A., & Newell, R. (2019). Frameworks and Models for Disseminating Curated Research Outcomes to the Public. *Sage Open*, *9*(2), 2158244019840112. https://doi.org/10.1177/2158244019840112

Cooper, M. M., Caballero, M. D., Carmel, J. H., Duffy, E. M., Ebert-May, D., Fata-Hartley, C. L., Herrington, D. G., Laverty, J. T., Nelson, P. C., Posey, L. A., Stoltzfus, J. R., Stowe, R. L., Sweeder, R. D., Tessmer, S., & Underwood, S. M. (2024). Beyond active learning: Using 3-Dimensional learning to create scientifically authentic, student-centered classrooms. *PLOS ONE*, *19*(5), e0295887. https://doi.org/10.1371/journal.pone.0295887

Employment in STEM occupations. (2024). U.S. BUREAU OF LABOR STATISTICS. https://www.bls.gov/emp/tables/stem-employment.htm

Feser, J., Vasaly, H., & Herrera, J. (2013). On the Edge of Mathematics and Biology Integration: Improving Quantitative Skills in Undergraduate Biology Education. *CBE—Life Sciences Education*, *12*(2), 124–128. https://doi.org/10.1187/cbe.13-03-0057

Figueroa, M. (2022). Podcasting past the paywall: How diverse media allows more equitable participation in linguistic science. *Annual Review of Applied Linguistics*, *42*, 40–46. https://doi.org/10.1017/S0267190521000118

Gross, L. J. (2000). Education for a Biocomplex Future. *Science*, *288*(5467), 807–807. https://doi.org/10.1126/science.288.5467.807

Hoffman, K., Leupen, S., Dowell, K., Kephart, K., & Leips, J. (2017). Development and Assessment of Modules to Integrate Quantitative Skills in Introductory Biology Courses. *CBE—Life Sciences Education*, *15*(2), ar14. https://doi.org/10.1187/cbe.15-09-0186

Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177. https://doi.org/10.1093/bfgp/elq001

Labov, J. B., Reid, A. H., & Yamamoto, K. R. (2010). Integrated Biology and Undergraduate Science Education: A New Biology Education for the Twenty-First Century? *CBE—Life Sciences Education*, *9*(1), 10–16. https://doi.org/10.1187/cbe.09-12-0092

McKeachie, W. J. (2014). Assessing, Testing, and Evaluating: Grading Is Not the Most Important Function. In *McKeachie's Teaching Tips* (pp. 73–84).

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education*, *30*(4), 159–167. https://doi.org/10.1152/advan.00053.2006

Occupational Outlook Handbook: Agricultural and Food Scientists. (2024). U.S. BUREAU OF LABOR STATISTICS. https://www.bls.gov/ooh/life-physical-and-social-science/agricultural-and-food-scientists.htm#:~:text=in%20May%202023.-,Job%20Outlook,on%20average%2C%20over%20the%20decade.

Schultheis, E. H., & Kjelvik, M. K. (2015). Data Nuggets: Bringing Real Data into the Classroom to Unearth Students' Quantitative & Inquiry Skills. *The American Biology Teacher*, *77*(1), 19–29. https://doi.org/10.1525/abt.2015.77.1.4

Schultheis, E. H., Kjelvik, M. K., Snowden, J., Mead, L., & Stuhlsatz, M. A. M. (2023). Effects of Data Nuggets on Student Interest in STEM Careers, Self-efficacy in Data Tasks, and Ability to Construct Scientific Explanations. *International Journal of Science and Mathematics Education*, *21*(4), 1339–1362. https://doi.org/10.1007/s10763-022-10295-1

Thompson, K. V., Nelson, K. C., Marbach-Ad, G., Keller, M., & Fagan, W. F. (2010). Online Interactive Teaching Modules Enhance Quantitative Proficiency of Introductory Biology Students. *CBE—Life Sciences Education*, *9*(3), 277–283. https://doi.org/10.1187/cbe.10-03-0028

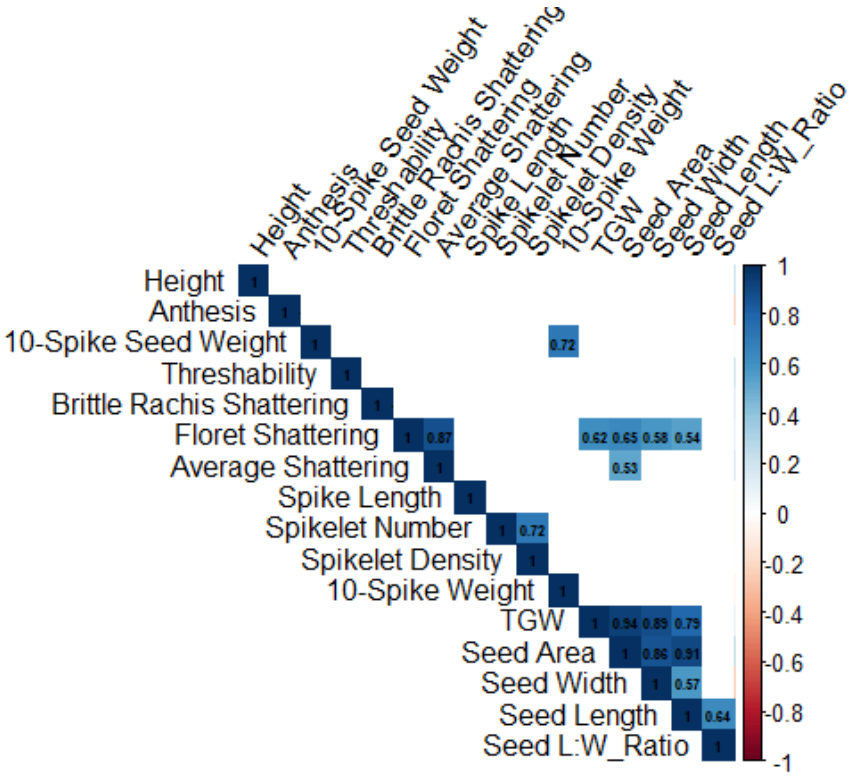
Wiggins, G. P., & McTighe, J. (2005). *Understanding by Design* (2nd edition). ASCD.

Appendices

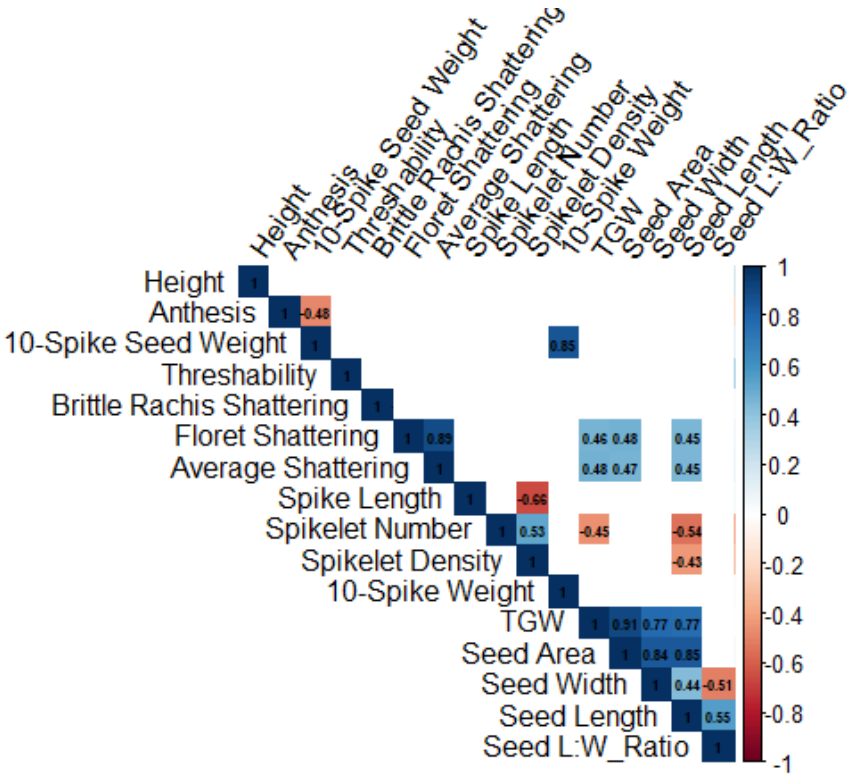
Chapter 1: Supplemental Material

Supplemental materials include Supplemental Figure S1a, S1b, S1c, and S1d, which are Pearson R Correlation Plots for traits significant at the $p < 0.05$ level, separated by Cycle. Figure S1a is Cycle 2, Figure S1b is Cycle 3, Figure S1c is Cycle 4, and Figure S1d is Cycle 5. Supplemental Table S1 includes the adjusted mean trait values for each cycle, calculated from mixed model analysis. Supplemental Table S2 depicts precipitation and temperature data for both locations and years of data collection.

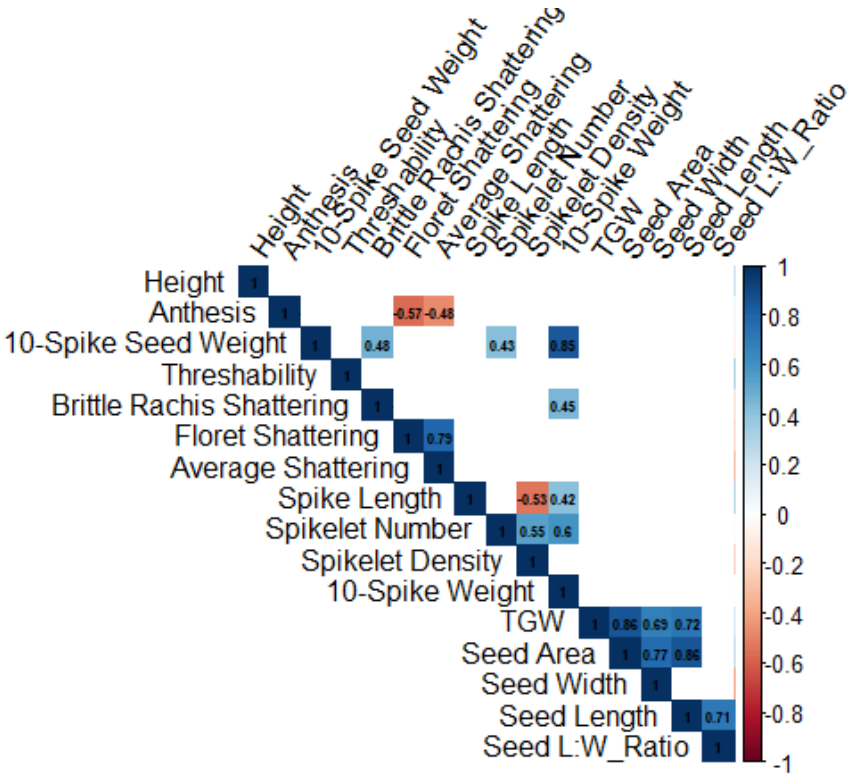
Supplemental Figure S1a: Cycle 2: Pearson R Correlation Coefficients for traits significant at the $p < 0.05$ level.



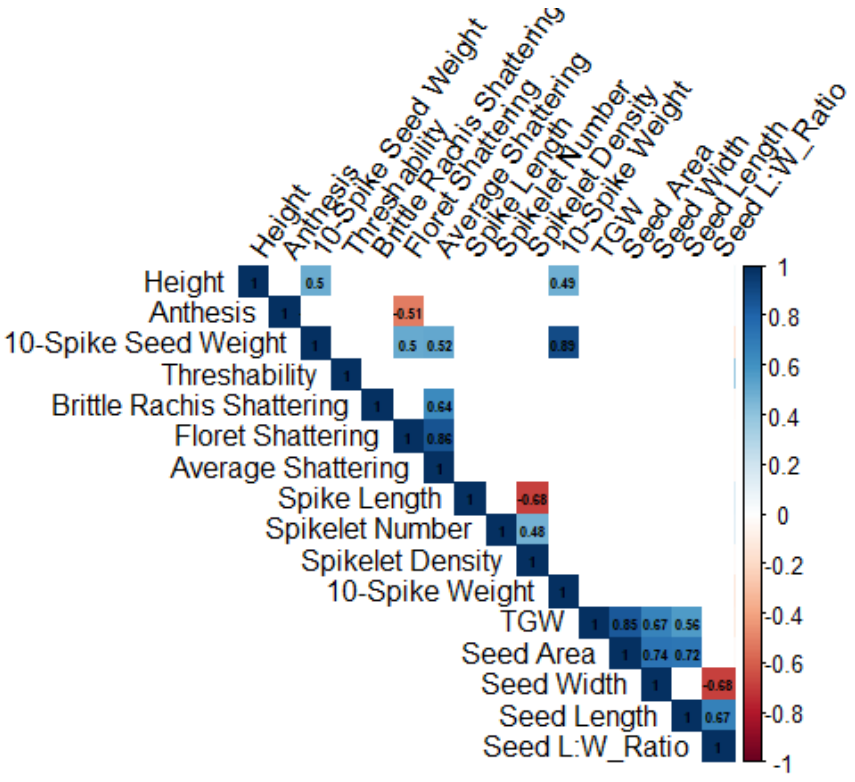
Supplemental Figure S1b: Cycle 3: Pearson R Correlation Coefficients for traits significant at the $p < 0.05$ level.



Supplemental Figure S1c: Cycle 4: Pearson R Correlation Coefficients for traits significant at the $p < 0.05$ level.



Supplemental Figure S1d: Cycle 5: Pearson R Correlation Coefficients for traits significant at the $p < 0.05$ level.



Supplemental Table S1: Adjusted mean trait values by Cycle, calculated from mixed model analysis.

Trait	Cycle 2	Cycle 3	Cycle 4	Cycle 5
Height (cm)	134.00	129.00	137.00	129.00
Anthesis (Julians)	173.00	174.00	173.00	174.00
10-Spike Seed Weight (g)	6.58	7.00	6.41	6.06
Threshability	4.87	4.96	4.66	5.53
Brittle Rachis Shattering	1.81	1.92	1.99	1.67
Floret Shattering	2.96	2.87	2.65	2.40
Average Shattering	2.39	2.40	2.32	2.04

Spike Length (cm)	27.90	28.10	27.90	27.20
Spikelet Number	22.00	22.10	22.50	22.10
Spikelet Density	0.80	0.80	0.81	0.82
10-Spike Weight (g)	12.80	13.30	12.50	12.30
TGW (g)	8.85	8.82	8.74	9.27
Seed Area (mm)	8.72	8.72	8.73	9.04
Seed Width (mm)	1.78	1.79	1.81	1.82
Seed Length (mm)	6.44	6.37	6.34	6.51
Seed L:W_Ratio (mm)	3.67	3.60	3.56	3.64

Supplemental Table S2: Cumulative Monthly and Average Precipitation and Temperature Data from May to August for the Twin Cities, MN and Lamberton, MN from 2000-2020, 2021, and 2022.

Year	Location	Precipitation				
		cm				
		May	June	July	August	Average (May-Aug)
2000-2020	Twin Cities, MN	10.39	12.04	9.58	11.15	10.80
2021	Twin Cities, MN	8.31	5.23	2.21	17.48	8.33
2022	Twin Cities, MN	8.46	2.87	3.00	10.85	6.29
2000-2020	Lamberton, MN	10.77	11.33	8.92	8.69	9.90
2021	Lamberton, MN	6.96	1.24	2.97	12.07	5.81
2022	Lamberton, MN	9.91	2.72	4.04	7.57	6.06
		Temperature				
		C				

		May	June	July	August	Average (May-Aug)
2000-2020	Twin Cities, MN	15.3	21.1	24.1	22.4	20.7
2021	Twin Cities, MN	15.7	24	24	23.3	22.1
2022	Twin Cities, MN	16.1	22.8	24.6	22.5	21.5
2000-2020	Lamberton, MN	14.2	20.4	22.6	20.7	19.7
2021	Lamberton, MN	14.3	22.8	22.5	21.9	20.3
2022	Lamberton, MN	14.6	21.8	23.2	20.9	19.9

Chapter 2: Supplemental Material

Supplemental Table S3: LD Decay calculated by chromosome.

Chromosome	LD Half Decay Distance
	(bp)
J01	993,543
J02	480,484
J03	943,103
J04	21
J05	843,678
J06	180
J07	132
S01	337,460
S02	68
S03	1
S04	623,798
S05	1,149,567
S06	599
S07	739,783
V01	3,266,042
V02	1
V03	1
V04	1
V05	1,035,475
V06	1,323
V07	1,085,225

Supplemental Table S4: Allele Frequencies for SNPs across each cycle found to be significant in GWAS. Major and minor allele counts and allele frequencies, and the number of homozygotes and heterozygotes are displayed within each cycle for each SNP.

Trait	SNP	Maj/Min Allele	Favorable Allele	Cycle	Maj Allele Count	Min Allele Count	Maj Allele Frequency	Min Allele Frequency	Favorable Allele Frequency	Homozygotes: Maj Allele	Homozygotes: Min Allele	Heterozygotes
10-spike Seed Weight	SJ05_21359377	A/T	A	2	66	10	0.868	0.132	0.868	30	2	6
				3	102	22	0.823	0.177	0.823	42	2	18
				4	94	40	0.701	0.299	0.701	31	4	32
				5	76	24	0.760	0.240	0.760	30	4	16
	SS05_56085353	G/C	C	2	40	36	0.526	0.474	0.474	13	11	14
				3	79	47	0.627	0.373	0.373	30	14	19
				4	89	45	0.664	0.336	0.336	34	12	21
				5	69	31	0.690	0.310	0.310	27	8	15
10-spike Weight	SJ02_133239962	C/T	C	2	41	35	0.539	0.461	0.539	8	5	25
				3	72	54	0.571	0.429	0.571	19	10	34
				4	58	76	0.433	0.567	0.433	10	19	38
				5	45	55	0.450	0.550	0.450	6	11	33
	SS05_161687303	C/T	C	2	62	14	0.816	0.184	0.816	24	0	14
				3	101	27	0.789	0.211	0.789	41	4	19
				4	96	36	0.727	0.273	0.727	33	3	30
				5	77	23	0.770	0.230	0.770	30	3	17
	SV02_81709239	G/A	G	2	62	14	0.816	0.184	0.816	24	0	14
				3	112	14	0.889	0.111	0.889	49	0	14
				4	108	26	0.806	0.194	0.806	43	2	22
				5	83	15	0.847	0.153	0.847	34	0	15
Average Shattering	SJ02_633868944	T/G	T	2	57	19	0.750	0.250	0.750	19	0	19
				3	101	25	0.802	0.198	0.802	40	2	21
				4	101	33	0.754	0.246	0.754	34	0	33
				5	84	16	0.840	0.160	0.840	34	0	16
	SS01_265896448	C/T	T	2	50	26	0.658	0.342	0.342	13	1	24
				3	82	42	0.661	0.339	0.339	23	3	36
				4	99	33	0.750	0.250	0.250	35	2	29
				5	64	36	0.640	0.360	0.360	14	0	36
	SS06_60087375	G/A	G	2	68	8	0.895	0.105	0.895	31	1	6
				3	106	18	0.855	0.145	0.855	45	1	16
				4	120	14	0.896	0.104	0.896	54	1	12
				5	93	7	0.930	0.070	0.930	43	0	7
	SV03_172019389	T/C	T	2	54	22	0.711	0.289	0.711	17	1	20
				3	86	40	0.683	0.317	0.683	27	4	32
				4	102	32	0.761	0.239	0.761	36	1	30
				5	67	33	0.670	0.330	0.670	21	4	25
Brittle Rachis	SJ03_171120774	A/C	C	2	69	7	0.908	0.092	0.092	33	2	3
				3	123	5	0.961	0.039	0.039	59	0	5
				4	128	6	0.955	0.045	0.045	61	0	6
				5	97	3	0.970	0.030	0.030	47	0	3
	SS02_462148698	C/G	C	2	57	19	0.750	0.250	0.750	23	4	11
				3	99	29	0.773	0.227	0.773	45	10	9
				4	86	48	0.642	0.358	0.642	40	21	6
				5	71	29	0.710	0.290	0.710	32	11	7
	SS03_37982539	A/C	A	2	55	21	0.724	0.276	0.724	22	5	11
				3	99	29	0.773	0.227	0.773	39	4	21
				4	101	33	0.754	0.246	0.754	41	7	19
				5	75	25	0.750	0.250	0.750	29	4	17
	SS07_96621872	A/G	G	2	71	5	0.934	0.066	0.066	33	0	5
				3	121	3	0.976	0.024	0.024	59	0	3
				4	128	6	0.955	0.045	0.045	61	0	6
				5	98	2	0.980	0.020	0.020	48	0	2
Floret Shattering	SV05_565531444	T/C	C	2	44	32	0.579	0.421	0.421	17	11	10
				3	63	59	0.516	0.484	0.484	24	22	15
				4	89	43	0.674	0.326	0.326	33	10	23
				5	71	29	0.710	0.290	0.290	30	9	11
Seed Area	SJ03_461795327	C/T	C	2	69	7	0.908	0.092	0.908	31	0	7
				3	118	8	0.937	0.063	0.937	55	0	8
				4	119	15	0.888	0.112	0.888	52	0	15
				5	92	8	0.920	0.080	0.920	42	0	8

	SJ04_50760398	C/T	T	2	47	29	0.618	0.382	0.382	9	0	29
				3	85	43	0.664	0.336	0.336	21	0	43
				4	98	36	0.731	0.269	0.269	31	0	36
				5	71	27	0.724	0.276	0.276	22	0	27
Seed Length	SJ03_461795327	C/T	C	2	69	7	0.908	0.092	0.908	31	0	7
				3	118	8	0.937	0.063	0.937	55	0	8
				4	119	15	0.888	0.112	0.888	52	0	15
				5	92	8	0.920	0.080	0.920	42	0	8
	SJ04_516190345	A/C	A	2	68	8	0.895	0.105	0.895	31	1	6
				3	112	16	0.875	0.125	0.875	48	0	16
				4	124	10	0.925	0.075	0.925	57	0	10
				5	97	3	0.970	0.030	0.970	47	0	3
	SJ05_484868935	T/C	T	2	41	35	0.539	0.461	0.539	10	7	21
				3	65	61	0.516	0.484	0.516	15	13	35
				4	66	64	0.508	0.492	0.508	22	21	22
				5	47	49	0.490	0.510	0.490	12	13	23
	SS01_375800564	A/G	G	2	66	10	0.868	0.132	0.132	28	0	10
				3	107	21	0.836	0.164	0.164	43	0	21
				4	106	26	0.803	0.197	0.197	40	0	26
				5	84	16	0.840	0.160	0.160	34	0	16
	SS03_357352971	G/A	A	2	70	6	0.921	0.079	0.079	32	0	6
				3	111	13	0.895	0.105	0.105	49	0	13
				4	116	18	0.866	0.134	0.134	50	1	16
				5	78	22	0.780	0.220	0.220	31	3	16
	SS05_323670052	T/C	T	2	41	35	0.539	0.461	0.539	12	9	17
				3	80	48	0.625	0.375	0.625	23	7	34
				4	72	60	0.545	0.455	0.545	18	12	36
				5	61	39	0.610	0.390	0.610	21	10	19
	SV02_106373759	G/A	A	2	59	17	0.776	0.224	0.224	21	0	17
				3	102	24	0.810	0.190	0.190	39	0	24
				4	110	22	0.833	0.167	0.167	44	0	22
				5	84	16	0.840	0.160	0.160	34	0	16
	SV07_486034824	A/G	G	2	71	5	0.934	0.066	0.066	33	0	5
				3	116	10	0.921	0.079	0.079	53	0	10
				4	129	5	0.963	0.037	0.037	62	0	5
				5	98	2	0.980	0.020	0.020	48	0	2
Spike-let Density	SJ07_436189945	G/A	G	2	73	3	0.961	0.039	0.961	35	0	3
				3	122	6	0.953	0.047	0.953	58	0	6
				4	129	5	0.963	0.037	0.963	62	0	5
				5	93	7	0.930	0.070	0.930	43	0	7
	SS02_373179566	A/G	G	2	59	17	0.776	0.224	0.224	21	0	17
				3	111	17	0.867	0.133	0.133	47	0	17
				4	108	26	0.806	0.194	0.194	41	0	26
				5	82	18	0.820	0.180	0.180	32	0	18
	SS02_465837109	T/G	G	2	63	13	0.829	0.171	0.171	25	0	13
				3	97	31	0.758	0.242	0.242	33	0	31
				4	101	33	0.754	0.246	0.246	37	3	27
				5	67	33	0.670	0.330	0.330	18	1	31
	SS05_30587500	G/A	G	2	64	12	0.842	0.158	0.842	26	0	12
				3	109	19	0.852	0.148	0.852	46	1	17
				4	115	19	0.858	0.142	0.858	48	0	19
				5	89	11	0.890	0.110	0.890	39	0	11
	SS07_82857819	G/A	G	2	68	8	0.895	0.105	0.895	30	0	8
				3	112	14	0.889	0.111	0.889	49	0	14
				4	126	6	0.955	0.045	0.955	60	0	6
				5	97	3	0.970	0.030	0.970	47	0	3
	SV02_457071045	C/A	C	2	39	37	0.513	0.487	0.513	2	1	35
				3	67	61	0.523	0.477	0.523	5	2	57
				4	74	60	0.552	0.448	0.552	8	1	58
				5	55	45	0.550	0.450	0.550	6	1	43
TGW	SJ03_28234197	G/A	G	2	47	29	0.618	0.382	0.618	9	0	29
				3	67	61	0.523	0.477	0.523	4	1	59
				4	78	56	0.582	0.418	0.582	11	0	56
				5	57	43	0.570	0.430	0.570	7	0	43
		C/T	C	2	40	36	0.526	0.474	0.526	4	2	32

	SJ06_540710206			3	69	59	0.539	0.461	0.539	6	1	57
				4	69	65	0.515	0.485	0.515	4	2	61
				5	53	47	0.530	0.470	0.530	3	0	47
	SS03_362322005	T/C	C	2	72	4	0.947	0.053	0.053	34	0	4
				3	112	14	0.889	0.111	0.111	50	1	12
				4	114	20	0.851	0.149	0.149	48	1	18
				5	73	27	0.730	0.270	0.270	27	4	19
Seed Width	SV01_409991737	G/A	G	2	53	23	0.697	0.303	0.697	18	3	17
				3	93	35	0.727	0.273	0.727	37	8	19
				4	108	24	0.818	0.182	0.818	45	3	18
				5	84	16	0.840	0.160	0.840	34	0	16

Supplemental Table S5: Significant BLASTP hits from comparing IWG version 3.0 to wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and rice (*Oryza* spp.) protein annotation. The top 10 protein matches by percent identity are displayed for each trait.

Trait	Description	Scientific Name	Taxid	Max Score	Total Score	Query Cover	E value	Percent Identity	Acc. Len	Accession
Seed Area	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	309	309	0.99	2e-106	83.12	249	XP_044336140.1
	zinc finger protein ZAT11-like [Triticum aestivum]	Triticum aestivum	4565	272	272	0.94	1e-92	79.82	206	XP_044354596.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	196	196	0.68	3e-63	79.5	160	XP_044345029.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	226	226	0.86	2e-74	77.27	204	XP_044337385.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	235	235	0.86	7e-78	76.38	202	XP_044353609.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	238	238	0.86	7e-79	76.24	208	XP_044345028.1
	zinc finger protein ZAT12-like [Hordeum vulgare]	Hordeum vulgare	4513	233	233	0.78	2e-77	72.78	200	KAE8789876.1
	hypothetical protein ZWY2020_047193 [Hordeum vulgare]	Hordeum vulgare	4513	234	234	0.78	3e-77	71.2	204	KAI5007245.1
	zinc finger protein ZAT12-like [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	232	232	0.78	1e-76	71.2	204	XP_044977776.1
	hypothetical protein CFC21_036857 [Triticum aestivum]	Triticum aestivum	4565	227	227	0.85	8e-75	70.44	200	KAF7024523.1
Brittle Rachis	uncharacterized protein LOC123077820 isoform X1 [Triticum aestivum]	Triticum aestivum	4565	2228	2228	0.99	0	97.58	1117	XP_044356084.1
	uncharacterized protein LOC123069300 isoform X1 [Triticum aestivum]	Triticum aestivum	4565	2203	2203	0.99	0	96.96	1116	XP_044348051.1
	uncharacterized protein LOC123443210 isoform X1 [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	2200	2200	0.99	0	96.87	1117	XP_044975445.1
	hypothetical protein D1007_20629 [Hordeum vulgare]	Hordeum vulgare	4513	2197	2197	0.99	0	96.78	1117	KAE8803499.1
	hypothetical protein ZWY2020_033255 [Hordeum vulgare]	Hordeum vulgare	4513	2194	2194	0.99	0	96.69	1117	KAI5006012.1
	predicted protein [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	2192	2192	0.99	0	96.6	1117	BAJ86599.1

	uncharacterized protein LOC123060738 isoform X1 [Triticum aestivum]	Triticum aestivum	4565	2183	2183	0.99	0	96.51	1112	XP_044339500.1
	hypothetical protein CFC21_034745 [Triticum aestivum]	Triticum aestivum	4565	2166	2166	0.99	0	96.06	1107	KAF7021861.1
	uncharacterized protein LOC123069300 isoform X2 [Triticum aestivum]	Triticum aestivum	4565	1957	1957	0.88	0	97.08	1008	XP_044348052.1
	uncharacterized protein LOC123443210 isoform X2 [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	1937	1937	0.88	0	96.48	1060	XP_044975446.1
Floret Shattering	CCR4-NOT transcription complex subunit 1-like [Triticum aestivum]	Triticum aestivum	4565	4791	4791	0.99	0	96.02	2405	XP_044377212.1
	hypothetical protein CFC21_063700 [Triticum aestivum]	Triticum aestivum	4565	4783	4783	0.99	0	95.93	2404	KAF7056277.1
	hypothetical protein CFC21_059713 [Triticum aestivum]	Triticum aestivum	4565	4793	4793	0.99	0	95.85	2408	KAF7051478.1
	CCR4-NOT transcription complex subunit 1-like [Triticum aestivum]	Triticum aestivum	4565	4788	4788	0.99	0	95.81	2408	XP_044371953.1
	CCR4-NOT transcription complex subunit 1-like [Triticum aestivum]	Triticum aestivum	4565	4772	4772	0.99	0	95.81	2404	XP_044384234.1
	predicted protein [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	645	645	0.13	0	95.6	318	BAJ93503.1
	CCR4-NOT transcription complex subunit 1-like [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	4612	4612	0.99	0	92.62	2421	XP_044983302.1
	CCR4-NOT transcription complex subunit 1-like [Hordeum vulgare]	Hordeum vulgare	4513	4604	4604	0.99	0	92.46	2425	KAE8818447.1
	CCR4-NOT transcription complex subunit 1-like [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	629	629	0.13	0	92.12	378	XP_044957186.1
	uncharacterized protein LOC123408030 [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	526	526	0.11	1e-167	92.06	422	XP_044957185.1
Spikelet Density	hypothetical protein CFC21_062851 [Triticum aestivum]	Triticum aestivum	4565	48.5	48.5	0.46	7e-08	85.71	131	KAF7055304.1
	ATP synthase alpha chain, mitochondrial, putative [Oryza sativa Japonica Group]	Oryza sativa Japonica Group	39947	47.8	47.8	0.43	7e-08	92.31	95	ABA97928.1
	hypothetical protein CFC21_082497 [Triticum aestivum]	Triticum aestivum	4565	46.2	46.2	0.45	1e-07	85.19	66	KAF7078018.1
	hypothetical protein CFC21_005278 [Triticum aestivum]	Triticum aestivum	4565	48.5	48.5	0.43	2e-07	92.31	224	KAF6987656.1
	hypothetical protein CFC21_095955 [Triticum aestivum]	Triticum aestivum	4565	48.5	48.5	0.43	2e-07	92.31	235	KAF7093548.1
	hypothetical protein CFC21_011662 [Triticum aestivum]	Triticum aestivum	4565	48.5	48.5	0.43	2e-07	92.31	528	KAF6995104.1

	ATP synthase F0 subunit 1 [Oryza sativa]	Oryza sativa	4530	48.5	48.5	0.43	2e-07	92.31	509	YP_010486804.1
	ATP synthase F0 subunit 1 [Oryza sativa]	Oryza sativa	4530	48.5	48.5	0.43	2e-07	92.31	509	QBE89896.1
	hypothetical protein DAI22_12g149301 [Oryza sativa Japonica Group]	Oryza sativa Japonica Group	39947	48.5	48.5	0.43	2e-07	92.31	509	KAF2908077.1
	hypothetical protein CFC21_019941 [Triticum aestivum]	Triticum aestivum	4565	48.5	48.5	0.43	2e-07	92.31	509	KAF7004760.1
Average Shattering	thioredoxin-like protein YLS8 [Triticum aestivum]	Triticum aestivum	4565	298	298	0.99	7e-105	99.3	142	XP_044325865.1
	thioredoxin-like protein YLS8 isoform X1 [Triticum aestivum]	Triticum aestivum	4565	298	298	0.99	8e-105	99.3	142	XP_044341806.1
	thioredoxin-like protein YLS8 [Triticum aestivum]	Triticum aestivum	4565	300	300	0.99	1e-104	99.3	220	XP_044368336.1
	thioredoxin-like protein YLS8 [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	296	296	0.99	5e-104	98.59	142	XP_044971294.1
	predicted protein [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	294	294	0.99	2e-103	97.89	142	BAJ88120.1
	hypothetical protein OsI_34094 [Oryza sativa Indica Group]	Oryza sativa Indica Group	39946	299	299	0.99	3e-103	96.48	274	EEC67208.1
	thioredoxin-like protein YLS8 isoform X1 [Oryza sativa Japonica Group]	Oryza sativa Japonica Group	39947	293	293	0.99	6e-103	96.48	142	XP_015612821.1
	thioredoxin-like protein YLS8 [Triticum aestivum]	Triticum aestivum	4565	293	293	0.99	8e-103	98.59	142	XP_044454246.1
	thioredoxin-like protein YLS8 [Oryza sativa Japonica Group]	Oryza sativa Japonica Group	39947	294	294	0.99	9e-103	96.48	183	NP_001408987.1
	hypothetical protein CFC21_032769 [Triticum aestivum]	Triticum aestivum	4565	266	266	0.99	3e-92	91.55	132	KAF7019609.1
Seed Length	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	309	309	0.99	2e-106	83.12	249	XP_044336140.1
	F-box protein At5g49610-like [Triticum aestivum]	Triticum aestivum	4565	690	690	0.96	0	82.13	427	XP_044451865.1
	hypothetical protein CFC21_013885 [Triticum aestivum]	Triticum aestivum	4565	683	683	0.96	0	81.89	425	KAF6997681.1
	F-box protein At5g49610-like [Triticum aestivum]	Triticum aestivum	4565	694	694	0.97	0	81.53	416	XP_044422117.1
	zinc finger protein ZAT11-like [Triticum aestivum]	Triticum aestivum	4565	272	272	0.94	1e-92	79.82	206	XP_044354596.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	196	196	0.68	3e-63	79.5	160	XP_044345029.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	226	226	0.86	2e-74	77.27	204	XP_044337385.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	235	235	0.86	7e-78	76.38	202	XP_044353609.1
	zinc finger protein ZAT12-like [Triticum aestivum]	Triticum aestivum	4565	238	238	0.86	7e-79	76.24	208	XP_044345028.1
	F-box protein At5g07610-like [Triticum aestivum]	Triticum aestivum	4565	289	289	0.53	1e-95	73.42	236	XP_044377567.1
TGW	proline-rich receptor-like protein kinase PERK8 [Triticum aestivum]	Triticum aestivum	4565	1304	1304	0.99	0	96.64	686	XP_044354783.1

	proline-rich receptor-like protein kinase PERK8 isoform X2 [Triticum aestivum]	Triticum aestivum	4565	957	957	0.67	0	96.98	463	XP_044346598.1
	proline-rich receptor-like protein kinase PERK8 isoform X1 [Triticum aestivum]	Triticum aestivum	4565	953	953	0.67	0	96.77	464	XP_044346597.1
	unnamed protein product [Triticum aestivum]	Triticum aestivum	4565	944	944	0.83	0	93.75	671	CDM80036.1
	proline-rich receptor-like protein kinase PERK8 [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	943	943	0.83	0	88.32	685	XP_044974313.1
	proline-rich receptor-like protein kinase PERK8 [Triticum aestivum]	Triticum aestivum	4565	943	943	0.68	0	96.35	687	XP_044338180.1
	hypothetical protein CFC21_038815 [Triticum aestivum]	Triticum aestivum	4565	942	942	0.83	0	93.75	685	KAF7026718.1
	hypothetical protein CFC21_038815 [Triticum aestivum]	Triticum aestivum	4565	937	937	0.83	0	93.59	686	KAF7026719.1
	hypothetical protein ZWY2020_031174 [Hordeum vulgare]	Hordeum vulgare	4513	918	918	0.83	0	87.99	686	KAI5003931.1
	proline-rich receptor-like protein kinase PERK9 [Hordeum vulgare]	Hordeum vulgare	4513	917	917	0.83	0	86.93	677	KAE8780019.1
10-spike Weight	E3 ubiquitin-protein ligase AIRP2-like [Triticum aestivum]	Triticum aestivum	4565	499	499	99%	0	99.59	242	XP_044322368.1
	E3 ubiquitin-protein ligase AIRP2-like [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	498	498	99%	0	99.17	242	XP_044964498.1
	E3 ubiquitin-protein ligase AIRP2-like [Triticum aestivum]	Triticum aestivum	4565	498	498	99%	0	99.17	242	XP_044455474.1
	hypothetical protein ZWY2020_009647 [Hordeum vulgare]	Hordeum vulgare	4513	486	486	97%	7e-175	98.73	316	KAI5008599.1
	hypothetical protein EE612_041186 [Oryza sativa]	Oryza sativa	4530	477	477	99%	2e-172	93.39	242	KAB8106720.1
	hypothetical protein DAI22_07g260900 [Oryza sativa Japonica Group]	Oryza sativa Japonica Group	39947	475	475	99%	1e-171	92.98	242	KAF2924318.1
	hypothetical protein CFC21_015619 [Triticum aestivum]	Triticum aestivum	4565	465	465	99%	9e-168	94.21	230	KAF6999615.1
	hypothetical protein OsI_27211 [Oryza sativa Indica Group]	Oryza sativa Indica Group	39946	463	463	99%	1e-166	86.26	262	EEC82625.1
	hypothetical protein CFC21_028315 [Triticum aestivum]	Triticum aestivum	4565	456	456	90%	3e-164	99.1	245	KAF7014298.1
	predicted protein [Hordeum vulgare subsp. vulgare]	Hordeum vulgare subsp. vulgare	112509	456	456	91%	6e-164	98.21	240	BAJ97797.1

Supplemental Table S6: IWG Chromosome naming translations from IWG V2.1 to IWG V3.1 genomes, including linkage groups.

V2.1	V3.1	Subgenome	Group
Chr01	J01	J	1
Chr06	J02	J	2
Chr07	J03	J	3
Chr10	J04	J	4
Chr14	J05	J	5
Chr18	J06	J	6
Chr20	J07	J	7
Chr02	S01	S	1
Chr04	S02	S	2
Chr08	S03	S	3
Chr11	S04	S	4
Chr13	S05	S	5
Chr17	S06	S	6
Chr21	S07	S	7
Chr03	V01	V	1
Chr05	V02	V	2
Chr09	V03	V	3
Chr12	V04	V	4
Chr15	V05	V	5
Chr16	V06	V	6
Chr19	V07	V	7

Chapter 3: Supplemental Material

Supplemental materials include links to the full Kernza® in Context Curriculum Google Drive, and the Breeding and Genetics Module. Additionally, the teacher's guide versions of the Data Nugget and Data Nugget Supplemental Activity are included.

Full Kernza® in Context Curriculum:

https://kernza.org/kernza-curriculum/

Kernza® in Context Breeding and Genetics Module:

https://drive.google.com/drive/folders/1XJQXNnlGs4ijBYRA0XpfRPDslfW8RlzU

DATA Nugget

A plant breeder's quest to improve perennial grain

Featured scientist: Hannah Stoll (she/her) from The University of Minnesota

Research Background:

Kernza® is a new grain crop that is similar to wheat. It can be ground into flour and used in bread, cookies, crackers and more! Unlike wheat, the rest of the plant can be eaten by livestock such as cattle. Another difference is that Kernza® is a **perennial**, meaning it grows in the ground for multiple years, whereas **annual** wheat only grows for one year. However, the challenge is that annual wheat makes more grain and is easier to harvest and sell. This means farmers currently prefer growing annual wheat over Kernza®.

Teacher Note: A perennial plant grows in the ground for multiple years, whereas an annual plant only grows for one year. Annuals invest a lot of their energy in above ground growth, which is one of the reasons that annual wheat produces more grain than a perennial crop. Perennial plants invest a lot of their energy below ground, and have much deeper roots. Perennial roots have several environmentally friendly characteristics. One benefit is preventing water contamination by soaking up extra water and nutrients from the soil. Also, the large perennial root systems hold soil in place year-round, preventing erosion.

One way to address this mismatch between annual and perennial crops is through **selective breeding**. This is when humans select individual plants with traits that are desirable for a specific reason. This group of individuals are strategically bred together. The breeder's goal is to shift the traits over generations. Scientists have only been working on breeding Kernza® for the past few decades; in comparison, humans started selecting annual wheat traits over 10,000 years ago! That is a lot of time to get the traits we are looking for.

Kernza® breeders are working on improving the same traits that have already been improved in annual wheat, including larger seed size. Kernza® scientists follow two main steps to breed plants 1) they select the best individuals from the population and 2) they intercross those individuals to create the next generation, or **breeding cycle**. With each breeding cycle, plant breeders see a slight improvement in the traits they selected.

Breeders can select plants based on phenotypes, genotypes, or both. Historically, plant breeders have selected based on desired **phenotypes**, or visible traits, only. Modern plant breeding can take advantage of the fact that we can now look at **genotypes**, or the genetic makeup, of individual plants quickly and at low costs. Scientists can use this

information to make quicker breeding improvements, so we don't have to wait another 10,000 years for high-yielding Kernza®!

Hannah is a scientist currently working on Kernza®. Hannah's passion for plant breeding was ignited during her high school years. She discovered the captivating world of genetics in her AP Biology class. It was then that she first realized the potential for breeding crop plants to make them more productive and viable for human consumption.

Hannah decided to join other scientists who work on Kernza® at the University of Minnesota. Here, scientists have completed four breeding cycles and are about to start the fifth. Hannah wanted to see whether different genetic makeups (genotypes) lead to differences in seed size (phenotypes). Her goal was to look at each plants' phenotype and genotype for seed size.

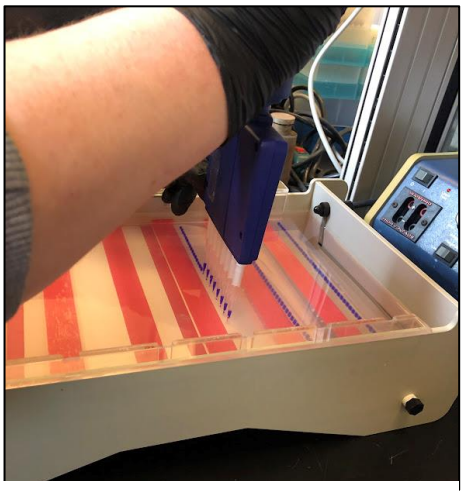
Hannah collecting data on Kernza®.

Teacher Note: You can take the time to connect this Data Nugget to course content related to genetics. These genes are transcribed from DNA to mRNA and translated from mRNA to protein (recall the Central Dogma of Biology). The resulting protein is what leads to the phenotype we see. You can use this opportunity to pause and watch videos or connect to your course content on transcription and translation if you desire.
Here are some potential videos:

- In depth (10 minutes) https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/transcription-and-rna-processing/v/transcription-and-mrna-processing
- Overview (3 minutes): https://www.youtube.com/watch?v=gG7uCskUOrA
- Hannah has also made a slideshow that you can use to go along with this activity. https://docs.google.com/presentation/d/1uxrR5UB3Vi1SlQrbej-u_rNC8BjC47bsj6XcS5PxnKE/edit?usp=sharing

To genotype a plant, scientists collect a small piece of leaf tissue, extract the DNA, and send the DNA to a lab for **sequencing**. This process tells scientists the genetic makeup that ultimately leads to the traits that we see. Specifically, sequencing data identifies **nucleotides**, or genetic building blocks of each plant's DNA. Plants have thousands of genes, which are made up of the DNA nucleotides A, T, C, and G.

Teacher Note: Phenotyping a single research plot costs at least $30 and requires more time to fully grow out the plant and replicate it in multiple environments to see how the phenotype differs by environment. Genotyping a single plant costs as low as $8 per sample (and prices continue to decrease as sequencing technologies improve).



DNA samples being tested for quality before genetic sequencing.

Sequencing data can be recorded in several ways. One common way is as **SNP data**, or **S**ingle **N**ucleotide **P**olymorphism data. You can think of SNP data as the recipe for proteins. In a SNP dataset, each SNP represents a *difference* in a **nucleotide.** Similar to using a different ingredient in a recipe, different nucleotides can result in a different phenotype.

By looking at SNP data, plant breeders can identify differences in genotypes that lead to certain phenotypes. Hannah started by evaluating 1,000 Kernza® plants from the first four breeding cycles. Data on phenotypes had already been recorded for these plants. Hannah then collected SNP data to determine their genotypes as well. She was looking for a pattern between genotypes and phenotypes. If she sees that different genotypes have different phenotypes, scientists can then rely on genotypes to select individuals to breed in future breeding cycles.

Teacher Note: For simplicity, students will see a representative snapshot of Hannah's data and just one seed-size trait, rather than working with thousands of datapoints! Moreover, there are 2 SNPs that are known to control seed area ("SNP 01" and "SNP 04") – those two SNPs combined make up the genotype for seed area.

Scientific Question: What genotype should a plant breeder select if they are trying to increase seed size in the next generation?

What is the hypothesis? Find the hypothesis in the Research Background and underline it. A hypothesis is a proposed explanation for an observation, which can then be tested with experimentation or other types of studies.

Scientific Data:

Complete Table 1a below, using the sequencing data for each plant to record its genotype.

Note: The table contains a subset, or smaller part, of data from Hannah's research. Hannah found that two nucleotides in the genotype were important for predicting the

phenotype of Kernza® seed size. Differences in SNP 01 and SNP 04 significantly impacted the phenotype values. Data for these SNPs are included below.

Table 1a. Phenotype and sequencing data for seed size from breeding cycles 1-4.

Plant ID	Breeding Cycle	**Seed area (mm)**	**SNP 01**	**SNP 04**	**Genotype**
1_1	1	5.4	A	G	AG
2_1	3	6.4	A	C	AC
3_1	2	7.5	T	A	TA
4_1	3	7.7	T	A	TA
5_1	2	6.9	A	C	AC
6_1	4	7.1	T	G	TG
7_1	1	6.6	A	C	AC
8_1	4	7.9	T	C	TC
9_1	3	6.8	A	C	AC
10_1	2	5.6	A	G	AG
11_1	1	6.1	A	A	AA
12_1	4	6.7	A	C	AC
13_1	1	6.7	T	G	TG
14_1	3	7.2	T	A	TA
15_1	4	7.6	T	C	TC
16_1	3	7.0	T	G	TG
17_1	1	6.5	A	C	AC
18_1	4	7.4	T	A	TA
19_1	2	7.3	T	A	TA
20_1	3	7.9	T	C	TC
21_1	3	7.6	T	A	TA
22_1	2	7.8	T	C	TC
23_1	1	7.0	T	G	TG
24_1	4	8.0	T	C	TC
25_1	2	6.3	A	A	AA
26_1	4	7.7	T	A	TA
27_1	1	5.7	A	G	AG
…	…	…	...	…	…
1000_1	2	6.9	T	G	TG

Now, complete Table 1b with the average seed area for each genotype.

Table 1b. Summary of seed area phenotype based on genotype for breeding cycles 1-4.

Genotype	**Average Seed Area (mm)**	Sample Size	Standard Deviation	Standard Error
AA	6.20	2	0.14	0.10
AC	6.65	6	0.19	0.08
AG	5.57	3	0.15	0.09
TA	7.49	7	0.20	0.07
TC	7.84	5	0.15	0.07
TG	6.94	5	0.15	0.07

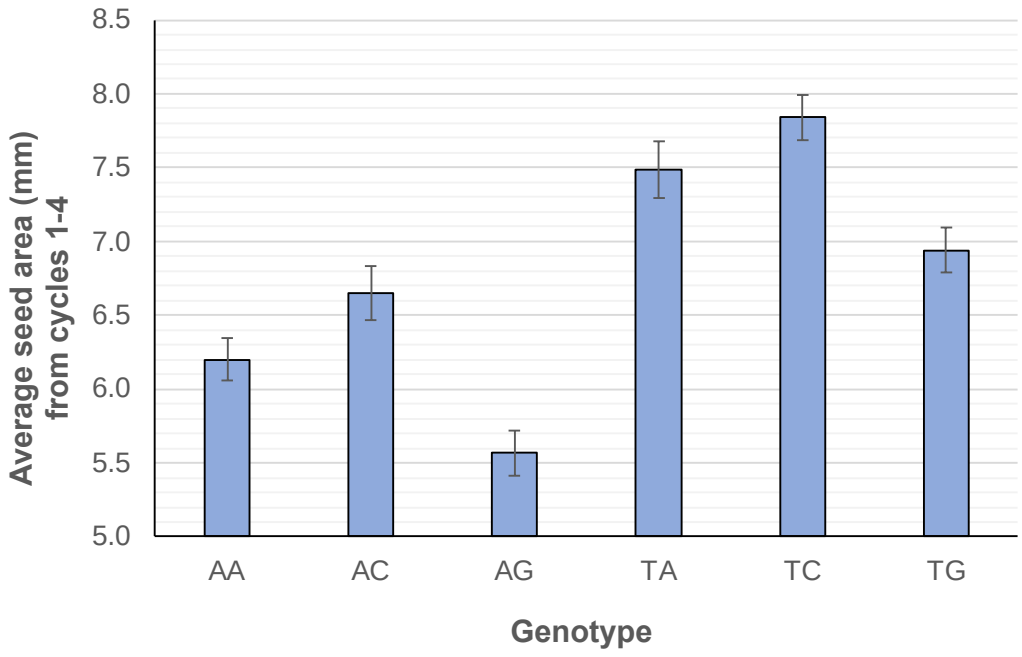
Next, use the data in Tables 1a and 1b to answer the scientific question.

What data will you graph to answer the question?

Independent variable: genotype

Dependent variable: average seed area (mm)- (phenotypic means for breeding cycles 1-4)

Draw your graph below: Identify any changes, trends, or differences you see in your graph. Draw arrows pointing out what you see and write one sentence describing what you see next to each arrow.



Interpret the data:

Make a claim that answers the scientific question, what genotype should a plant breeder select if they are trying to increase seed size in the next generation?

Genotype "TC" leads to the largest seed area phenotype in cycles 1-4. Plant breeders should select for this genotype for future generations of Kernza to increase seed size.

What evidence was used to write your claim? Reference specific parts of the tables or graph.

TC genotype has an average seed area of 7.84 for breeding cycles 1-4. The second largest average seed area is TA, at 7.49 mm. The lowest average seed area was AG, at 5.57 mm. The standard error is 0.10 mm and the standard deviation is 0.14 mm.

Teacher Note - Error Bars: You can have students add error bars to their graphs to deepen this discussion or remove SE or SD from the table for younger students. Standard deviation (SD) is the most common measure of variation for normally distributed data. It is a measure of the average distance of all values from their mean. The smaller the bars, the less variation around the mean. Standard error (SE) is the SD divided by the square root of the study's sample size ($SE=SD/\sqrt{n}$). Unlike SD, SE reflects uncertainty in our estimate of the mean. The larger our sample size and the less variation in the data, the more confident we can be in our estimate of the mean. Upper error bars are calculated by adding one SE or SD to the mean, and lower bars are calculated by subtracting one SE or SD from the mean.

The simplest measure of spread or variation in a data set is the range, which is the difference between the largest and smallest values in the data set. For students unfamiliar with SD or SE, a discussion of range can help bring their attention to not only central tendency in the data, but also variation around the mean. Be sure to note with students that though range is easy to calculate, it can be misleading; one outlier can make it appear the

Explain your reasoning and why the evidence supports your claim. Connect the data back to what you learned about how DNA sequencing for genotypes can inform selective breeding.

By combining DNA sequencing data with phenotype data that was measured directly from plants, we can find the genotypes that have the desired traits. We can see that the average seed area

phenotypic value for each genotype is different for each genotype, suggesting the genotype controls the phenotype.

In this case, Hannah was looking to see which genotypes had the largest seed size area. The genotype "TC" leads to a seed area phenotype that is over 2 mm larger than the average seed area for individuals with the "AG" genotype. If plant breeders selected TC to breed instead of individual plants with the AG phenotype, we can expect to see the seed size of Kernza plants to increase over generations.

Did the data support Hannah's hypothesis? Use evidence to explain why or why not. If you feel the data are inconclusive, explain why.

Yes, we can see that the average seed area phenotypic value for each genotype is different for each genotype, suggesting the genotype controls the phenotype. However, one could argue that there may be more SNPs at play that are controlling the phenotype.

Teacher Note: In fact, this is most likely true – most traits are controlled by dozens if not hundreds of genes. For example, maybe there are really 3 SNPs controlling seed area, so we should be looking at genotypes like "ACC", "TGA", etc. If there really were 3 SNPs controlling the phenotype, we would see greater resolution of differences between phenotypes, and there would be more combinations of possible genotypes and phenotypes.

Students may also argue that you can't definitively say that phenotypic means for each genotype are significantly different – the standard deviations help you determine that several are, but a T-test between each genotype pair (AA vs. AC, AA vs. AG, etc) would be needed

Your next steps as a scientist: Science is an ongoing process. What new question(s) should be investigated to build on Hannah's research? How do your questions build on the research that has already been done?

See the following Teacher Note.

Teacher Note: Students may wonder how the process of selecting based on genotype from predicted phenotypes works for traits other than seed area? In the future, Hannah could test this for other traits, such as grain yield and seed shattering.

Teacher Note: Interested in getting more practice applying the concepts of genotypic selection? Want to dive deeper into how Hannah and other plant breeders use genotype to predict trait values for the next generation?

If so, continue to the **supplemental activity** for more practice. In the supplemental activity, students will test a hypothesis to see if plant breeders can predict the phenotype values of un-phenotyped plants using genotypic data and validate their hypothesis using real data. This supplement gives students a view into how scientists apply these techniques in a real breeding program, and how both genotype and phenotype information leveraged to make quicker progress in future breeding cycles.

DATA Nugget

Supplemental Activity: A plant breeder's quest to improve perennial grain

Featured scientist: Hannah Stoll (she/her) from The University of Minnesota

In the previous activity, you used data from Breeding Cycles 1-4 to look at whether different genotypes had different phenotypes and found that certain genotypes were associated with larger seeds. You will now explore how Hannah used genotype and phenotype data from those previous generations to make predictions about Breeding Cycle 5!

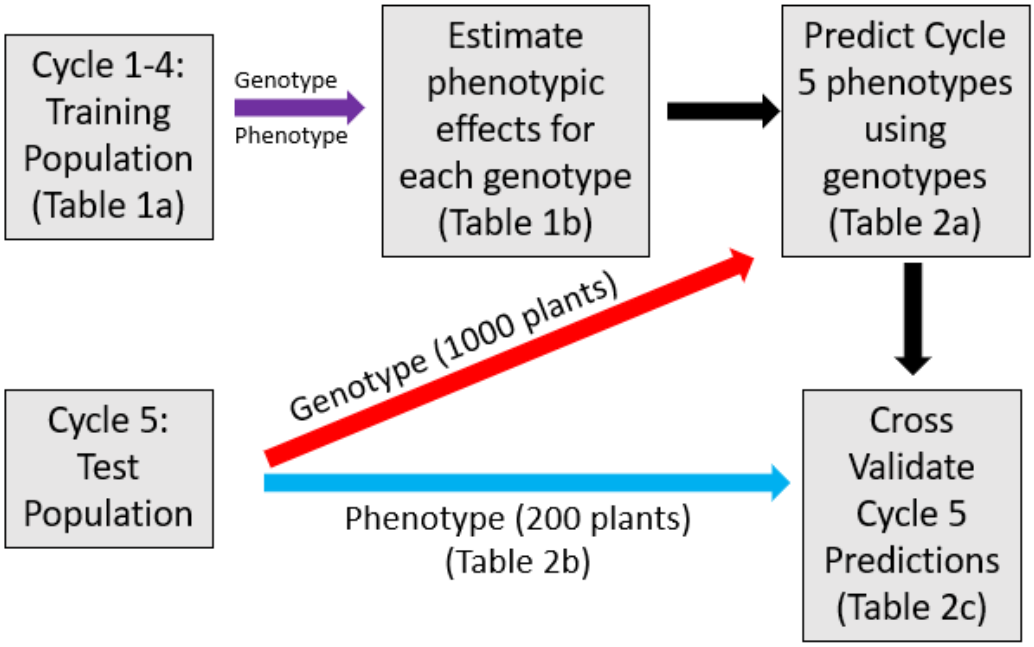
Supplemental Research Background:

Breeders are interested in increasing seed size in Kernza®, and have done so up to this point selecting individuals based on phenotype, not genotype. This is an issue, since phenotyping plants requires scientists to grow thousands of plants, which is time-consuming and expensive. Hannah wants to know if breeders can instead rely on genotype to select the individuals that will contribute to future breeding cycles.

Hannah set out to evaluate whether she can accurately predict the phenotype of an individual plant from their genotype. She went back to data from Breeding Cycles 1-4. She used the average phenotypes for each genotype to make predictions about a new generation, Breeding Cycle 5. If her predictions were accurate, she would know that she could use genotype to predict phenotype!

Hannah looked at the resources and funds she had to collect data on Breeding Cycle 5. She saw that she had funds to genotype all 1,000 plants but could only phenotype a subset of 200 plants.

Hannah compared phenotyped individuals from Cycle 1-4 to the 200 phenotyped individuals from Cycle 5. If the predicted phenotype values from Cycle 1-4 for each genotype are accurate, then those values can be used to predict the phenotypes of the remaining 800 plants in Cycle 5, which she didn't have the resources to grow and obtain phenotypes directly.



Schematic of how a plant breeder may use both phenotypic and genotypic data in a Kernza® breeding program to make selections for future generations. Table 1a-1b reference the tables from the Data Nugget.

Scientific Question: How well does genotypic data from Breeding Cycles 1-4 predict phenotypes of Breeding Cycle 5 individuals?

Scientific Data:

Use the data below to answer the scientific questions:

Remember: Hannah doesn't have the resources to phenotype all 1000 of her Cycle 5 plants, but she does have the resources to genotype all of them. ***Use the phenotypic values you calculated in the first activity (Table 1b) to predict seed area for this subset of Cycle 5 plants****.*

Table 2a. Test Data: Fill in the genotypes and the associated predicted Cycle 5 phenotype.

Plant ID	Breeding Cycle	**Seed Area (mm)**	**SNP 01**	**SNP 04**	**Genotype**
1_2	5	5.57	A	G	AG
2_2	5	6.20	A	A	AA
3_2	5	7.49	T	A	TA
4_2	5	7.49	T	A	TA
5_2	5	6.20	A	A	AA
6_2	5	6.94	T	G	TG
7_2	5	6.65	A	C	AC
8_2	5	7.84	T	C	TC
9_2	5	6.65	A	C	AC
10_2	5	5.57	A	G	AG
11_2	5	6.20	A	A	AA
12_2	5	6.65	A	C	AC
13_2	5	6.94	T	G	TG
14_2	5	7.49	T	A	TA
15_2	5	7.84	T	C	TC
16_2	5	6.94	T	G	TG
17_2	5	5.57	A	G	AG
18_2	5	7.49	T	A	TA
19_2	5	7.49	T	A	TA
…	…	…	…	…	…
1000_2	5	7.84	T	C	TC

Hannah phenotyped 200 of the 1000 plants in the Cycle 5 population and calculated the following phenotypic values.

Table 2b. Test data calculated from phenotype and genotypes of 200 Cycle 5 plants.

Genotype	**Average Seed Area (mm)**	Sample Size	Standard Deviation	Standard Error
AA	6.14	30	0.01	0.08
AC	6.73	33	0.03	0.17
AG	5.53	28	0.02	0.10
TA	7.40	35	0.02	0.10
TC	7.78	41	0.02	0.15
TG	6.92	33	0.01	0.08

Now, calculate the difference between your predicted phenotypes (from Cycle 1-4 data) and the actual phenotypes from Cycle 5 plants.

Table 2c. Validation table.

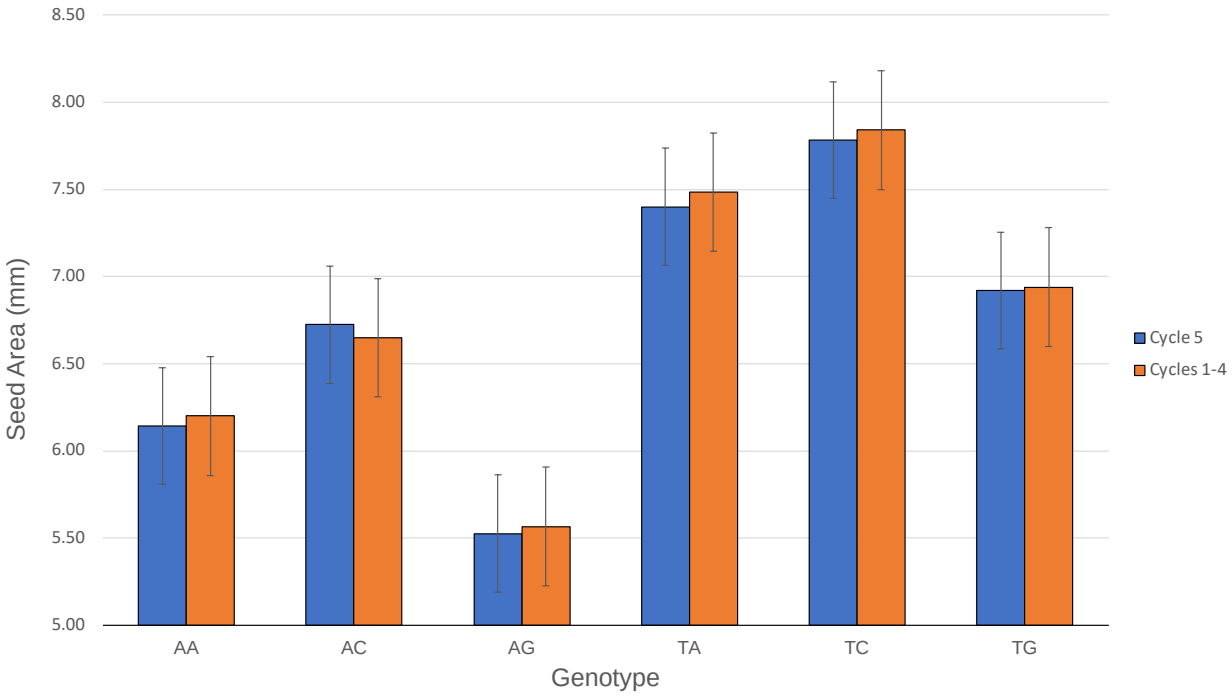
Genotype	**Cycle 1-4: Average Seed Area (mm)**	Cycle 1-4: Standard Deviation	**Cycle 5: Average Seed Area (mm)**	Cycle 5: Standard Deviation	**Difference** between Cycle 1-4 Average and Cycle 5 Average
AA	6.20	0.14	6.13	0.08	0.07
AC	6.65	0.19	6.73	0.17	-0.08
AG	5.57	0.15	5.53	0.10	0.04
TA	7.49	0.20	7.40	0.10	0.09
TC	7.84	0.15	7.78	0.15	0.06
TG	6.94	0.15	6.92	0.08	0.02

What data will you graph to answer the question?

Independent variable: genotype (for cycle 1-4 data and Cycle 5 data)

Dependent variable: average seed area (mm) for cycle 1-4 and cycle 5

Draw your graph below: Identify any changes, trends, or differences you see in your graph. Draw arrows pointing out what you see and write one sentence describing what you see next to each arrow.



Interpret the data:

Make a claim that answers the scientific question, how well does genotypic data from Breeding Cycles 1-4 predict phenotypes of Breeding Cycle 5 individuals?

The genotypic data from Breeding Cycles 1-4 accurately predicted phenotypes of Breeding Cycle 5.

What evidence was used to write your claim? Reference specific parts of the table or graph.

The phenotypic predictions were anywhere from -.08 to +.09 different between Cycle 1-4 and Cycle 5.

In Table 2c, the differences between Cycle 1-4 and Cycle 5 phenotyped individuals (last column) range between 0.02-0.09 (absolute value) between the 2 sets of data. Students could circle the last column in Table 2b. Additionally, if they graph the data as a clustered bar plot, students can see the phenotypic estimates for Cycle 1-4 and Cycle 5 are basically the

same, meaning Cycle 1-4 phenotypic data can be used to predict Cycle 5 data.

Explain your reasoning and why the evidence supports your claim. Connect the data back to what you learned about how genotyping information can be leveraged in a Kernza® breeding program.

We showed that the Cycle 1-4 phenotyped and genotyped individuals' data could be leveraged to predict Cycle 5 phenotypic values. The difference between Cycle 1-4 data and Cycle 5 cross validation data was very small, suggesting they are very similar, and genotypic data can be used to predict phenotypes with high confidence.

The estimates were not identical, but similar enough to observationally conclude they were not significantly different. However, since we didn't do a mean comparison test (T-test), one could argue that you can't definitively say that Cycle 1-4 and Cycle 5 means are not significantly different. However, observationally, they are nearly the same.

This difference is not greater than any of the respective standard deviations listed. This means Hannah and other scientists can use genotypic data to predict phenotypes.

Note, the appropriate way to test this hypothesis would be with a T-test to compare means between Cycle 1-4 and Cycle 5, but a simple look at the data suggests they are not very different from one another.

Teacher Note: A T-test is a statistical test used to determine if there is a *significant* difference between the means of two groups. In this activity, we did not perform a T-test (we simply looked at the means for the two groups and determined that they appeared similar), but a T-test would give scientists a definitive, statistically significant conclusion.

Your next steps as a scientist: Science is an ongoing process. What new question(s) should be investigated to build on Hannah's research? How do your questions build on the research that has already been done?

Next, Hannah will use the validated prediction data to make selections from the Cycle 5 plants – she will choose the plants that have the largest predicted seed area for the next breeding generation – Cycle 6.

What is the minimum number of plants a scientist needs to phenotype and genotype to predict a phenotype using the genotype for a trait? This probably differs for a given trait, but testing this would allow Hannah to optimize the number of plants she phenotypes… potentially saving her lots of time and resources!