Investigating genomic prediction in Ontarian barley varieties

Student Name	**:**	Abelhard Moshe Jaime Jauwena
Student Number	**:**	1040413
Course Name	**:**	Bioinformatics Master's Project
Course Number	**:**	BINF*6999
Submission Date	**:**	December 15, 2023
University Advisors	**:**	
Informatics Expertise	:	Dr. Yan Yan, School of Computer Science
Biological Expertise	:	Dr. Lewis Lukens, Department of Plant Agriculture

Table of Contents

Abstract

Increasing food production is important in Ontario, Canada, and one of the ways to accomplish this is to increase barley production. The historical yield trend of Ontarian barley varieties from 1958 to 2021 was investigated. Genome-wide allelic differences in these varieties were also examined to see if they correlate with yield differences over time. Lastly, the utility of genomic prediction in estimating an unknown variety's performance using these genome-wide allelic differences was also investigated.

Ontarian barley yields increased steadily from 1958 to 2021 at a rate of 33.7 kg ha^{-1} $year^{-1}$, with equal contributions from both genetic and environmental factors. Yield increases over time also correspond with genome-wide allelic shifts in these varieties. Furthermore, the predicted yields obtained using genomic prediction correlated moderately with observed yields, with a mean prediction accuracy of 0.557. Therefore, genomic prediction has the potential to accelerate barley production in the future.

Introduction

Barley (*Hordeum vulgare L.*) is one of the main cereal crops grown in Ontario, Canada (Ontario Ministry of Agriculture, 2023). Therefore, increasing barley production can significantly improve food production in Ontario. It is helpful to investigate the historical yields of Ontarian barley varieties to identify if they have been increasing and at what rate (So et al., 2022). Doing so also allows one to elucidate the extent to which genetic and environmental factors influence the magnitude and direction of yield trends over time (So et al., 2022). Here, genetic factors are factors related to the breeding of genetically superior barley varieties (Piepho et al., 2014).

Conversely, environmental factors are any factors unrelated to breeding efforts, including improved agronomic practices (e.g., improved farm machinery or more effective weed, disease, and pest control) and climatic changes (Piepho et al., 2014).

Genetic improvement, which is achieved via breeding genetically superior varieties, is a key component in accelerating food production (Bailey-Serres et al., 2019). Therefore, the genetic factors that have contributed to past barley performances in Ontario can be used to accelerate future progress. One tool that can help achieve this aim is genomic prediction, which involves training a statistical model on a genotyped and phenotyped population to predict the phenotype of genotyped individuals (Zhao et al., 2023). Specifically, genomic prediction can provide accurate estimates of crop performance based on their genotype and has been tested on different grains, such as rice (Bartholomé et al., 2022) and winter wheat (Jackson et al., 2023; So et al., 2022).

The objectives of this project are threefold. First, it aims to elucidate the historical yield trend of Ontarian barley varieties and whether it is primarily influenced by genetic or environmental factors. Second, it aims to identify genome-wide allelic differences in these barley varieties and whether they correlate with yield differences over time. Third, it aims to investigate whether genomic prediction can estimate an unknown variety's performance using these genome-wide allelic differences.

Materials and Methods

Barley performance trial dataset and elucidating historical yield trends

A dataset containing the historical yield entries for 372 Ontarian barley varieties, recorded from 1958 to 2021, was obtained. This dataset was generated by the Ontario Cereal Crop Committee (OCCC) via their annual barley performance trials, which assess agronomic traits in barley varieties and provide data for variety testing (So et al., 2022; Yang et al., 2023). Entries were recorded from 1958 to 2021 in 199 locations across six areas. The dataset was filtered using two criteria. First, entries with missing yields were removed, as they cannot be used in analyses. Second, varieties that appeared less than twice cumulatively from 1958 to 2021 were also removed. The cutoff value was set to two to preserve as many entries as possible while ensuring they remain connected. The year of the first entry into trials (YFE) was also added for each entry. 13,541 historical yield entries from 215 varieties remained after filtering, and these entries were still obtained from 199 locations across six areas.

The overall historical barley yield trend was obtained by linearly regressing yields by years. Adjusted yield means were then obtained for each year using the “emmeans” function from the emmeans R package (Lenth et al., 2023). Variance components (**Tab. 1**) were estimated by fitting a three-way model from Laidig et al. (2008) to the dataset using the “lmer” function from the “lme4” R package (Bates et al., 2023). The model is given as follows:

$$y_{ijk} = \mu + G_i + L_j + Y_k + (LY)_{jk} + (GL)_{ij} + (GY)_{ik} + (GLY)_{ijk} \quad (1)$$

where y_{ijk} is the mean yield of the i^{th} variety in the j^{th} location and k^{th} year, μ is the mean overall yield, G_i is the main effect of the i^{th} variety, L_j is the main effect of the j^{th} location, Y_k is the main effect of the k^{th} year, $(LY)_{jk}$ is the jk^{th} location × year interaction effect, $(GL)_{ij}$ is the ij^{th} variety × location interaction effect, $(GY)_{ik}$ is the ik^{th} variety × year interaction effect, and $(GLY)_{ijk}$ is a residual that encompasses the variety × location × year interaction effect and the error of the mean (Piepho et al., 2014). All terms were treated as random effects (So et al., 2022).

To model the genetic trend, defined as the trend resulting from the breeding of genetically superior varieties (Piepho et al., 2014), the following model was fitted:

$$G_i = \beta r_i + H_i \qquad (2)$$

where β is a fixed regression coefficient for the genetic trend, r_i is the YFE, and H_i is the random deviation of G_i from the genetic trend line, which is assumed to follow a normal distribution with a mean of zero and a variance of σ^2_H (Piepho et al., 2014). β can be used to quantify the genetic trend, as it gives the yield increase per year measured in the same units as y_{ijk} (Piepho et al., 2014; So, n.d.).

Modeling the environmental trend was done by fitting the following model:

$$Y_k = \gamma t_k + Z_k \qquad (4)$$

where γ is a fixed regression coefficient for the environmental trend, t_k is a continuous covariate for the calendar year, and Z_k is a random residual following a normal distribution with a mean of zero and a variance of σ^2_H (Piepho et al., 2014). γ can be used to quantify the environmental trend, as it also gives the yield increase per year, measured in the same units as

y_{ijk} (Piepho et al., 2014; So, n.d.). Adjusted yield means were also obtained for each year for the genetic and environmental trend.

GBS dataset and variant calling

Paired-end reads were obtained via genotyping-by-sequencing (GBS), which consisted of several steps. DNA samples were first extracted from Ontarian barley varieties in May 2022 and June 9, 2023. These samples were digested by a rare-cutting restriction enzyme, *PstI* (CTGCAG), and a common-cutting restriction enzyme, *MspI* (CCGG), resulting in DNA fragments (Poland et al., 2012). A set of barcoded adapters was ligated to these fragments for identification purposes, and these ligated fragments were then pooled and amplified using PCR (Poland et al., 2012). The resultant GBS library was then put on 96-well plates and subjected to high-throughput sequencing (HTS) using an Illumina HiSeq4000, producing paired-end reads (So et al., 2022).

Reads were processed using "Fast-GBS," a highly accurate genotyping pipeline (**Fig. 1**) (Torkamaneh et al., 2017). First, "Sabre" (ver. 1.00) (najoshi, 2013) was used to remove barcodes from these reads and demultiplex them (Torkamaneh et al., 2017). Second, "Cutadapt" (ver. 3.10.2) (Martin, 2011) was used to trim Illumina adapters and repeated bases from the 5' end of both the forward and reverse reads. The adapter sequences were "AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC" the forward reads and "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT" for the reverse reads. "FastQC" (ver. 0.11.9) (Andrews, 2010) was also used to check the quality of both the demultiplexed reads and the demultiplexed and trimmed reads. Third, "BWA" (ver. 0.7.17) (Li & Durbin, 2010) was used to

build the index for the “Morex V3” assembly (ver. 110.4) (European Nucleotide Archive, n.d.), which was used as the reference genome. The demultiplexed and trimmed reads were then aligned to this index, converted into BAM format using “SAMtools” (ver. 1.17) (Li, 2011), and sorted. SAMtools was also used to build another index for the reference genome. This index was used for calling single nucleotide polymorphisms (SNPs) on the sorted BAM files with “Platypus” (ver. 0.8.1) (Rimmer et al., 2014). Platypus was specified to call SNPs on reads with a mapping quality score of at least 20 and a base quality score of at least 20 (Rimmer, 2017). Furthermore, Platypus only called SNPs with at least two reads per locus and did not call indels (Rimmer, 2017). Variant calling results were outputted as VCF files. Lastly, “VCFtools” (ver. 0.1.16) (Danecek et al., 2011) was used to process the called SNPs. VCFtools was specified to remove indels and sites not flagged as “PASS” by Platypus for not meeting the specified quality thresholds (Auton & Marcketta, 2015).

Fig. 1: A diagram of the Fast-GBS pipeline. The main steps are outlined at the center, the tools used for each step are on the left-hand side, and the input and output files are on the right-hand side. Solid arrows indicate the main steps of the pipeline, while dashed arrows indicate the tools and the input and output files.

The called SNPs were filtered using TASSEL (ver. 5.2.90) (Bradbury et al., 2007) and VCFtools (Danecek et al., 2011). Varieties with more than 90% missing loci along with highly heterozygous varieties were removed (Smith, 2023). Loci with more than 80% missing data across varieties and alleles with minor allele frequencies less than or equal to 5% were also removed (Smith, 2023). Beagle (ver. 5.4) (Browning et al., 2018) was used to impute missing loci. The "snpgdsGetGeno" function from the "SNPRelate" R package (Zheng, 2018) was used to

check for the number of SNPs identified after filtering. 35,481 SNPs belonging to 116 varieties were identified.

Identifying genome-wide allelic differences using PCA

A total of 54 overlapping varieties, defined as varieties that have both yield and SNP data, were obtained. Varieties were categorized as “old” or “new” based on whether they were first introduced pre- or post-1990, respectively. Varieties were also categorized as a six-row or two-row variety by referencing data from the Government of Canada database (Government of Canada, 2022). Finally, the adjusted yield means for each variety was obtained using emmeans (Lenth et al., 2023).

SNPs for these overlapping varieties were pruned with a linkage disequilibrium (LD) threshold of 0.1 using the “snpgdsLDpruning” function from SNPRelate (Zheng, 2018). 7,813 SNPs remained after filtering and pruning. Principal component analysis (PCA) was conducted on these SNPs using the “snpgdsPCA” function (Zheng, 2018).

Performing genomic prediction

Using VCFtools (Danecek et al., 2011), SNPs for the overlapping varieties were converted to the “012” format, where 0 represents a homozygous genotype for the reference allele, 1 represents a heterozygous genotype, and 2 represents a homozygous genotype for the non-reference allele (Auton & Marcketta, 2015; Strandén & Christensen, 2011). Afterward, these SNPs were subtracted by one and converted to the “-101” format for compatibility with the

"rrBLUP" R package (Endelman, 2023). SNP data for the overlapping varieties was joined with the corresponding data for adjusted yield means and YFEs.

Cross-validation was used to build a genomic prediction model that predicted the adjusted yield means of the overlapping varieties based on their genotype. Specifically, the "mixed.solve" function from rrBLUP (Endelman, 2023) was used to capture the relationship between SNPs and adjusted yield means for a subset of the varieties. Then, the estimated marker effects were used to predict the adjusted yield means of varieties outside the subset. 150 iterations of 5-fold cross-validation were performed to obtain a more accurate estimate of the performance of the genomic prediction model (Tougui et al., 2021). Model prediction accuracy (PA) per fold was calculated by obtaining the correlation between predicted and observed adjusted yield means. The mean PA per iteration was calculated by averaging the PA across all five folds in a given iteration. The mean overall PA was calculated by averaging the mean PA per iteration across all 150 iterations. Additionally, marker effects for old varieties were obtained using mixed.solve (Endelman, 2023) and used to predict the yields of new varieties. Marker effects for new varieties were also obtained and used to predict the yields of old varieties.

Results

Ontarian barley yields have historically increased, and the location-by-year interaction effect accounted for most of the variation in yields

Historical barley yields increased steadily over the years, with the rate of yield increase estimated at 33.7 kg ha^{-1} year^{-1} ($P < 0.001$) (**Fig. 2**). The average adjusted yield means was 3196.896 kg ha^{-1} from 1958 to 1989, whereas it was 4461.418 kg ha^{-1} from 1990 to 2021.

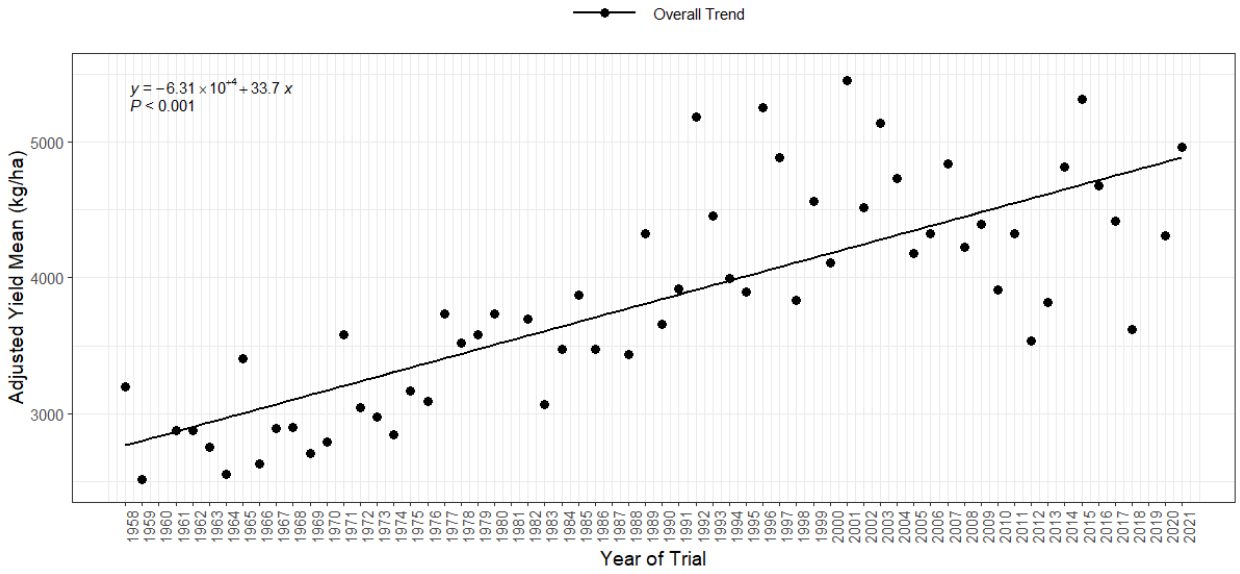

Fig. 2: Overall trend for historical barley yields, obtained by plotting the adjusted yield means against the year of trial.

The location-by-year interaction accounted for most, 57%, of yield variation (**Tab. 1**). Furthermore, the individual effects of location and year also accounted for 14.6% and 10.5% of yield variation, respectively (**Tab. 1**). On the other hand, the effect of varieties on yields was stable, as yields only accounted for 2.9% of yield variation (**Tab. 1**). The variety-by-year and

variety-by-location interactions also only accounted for 0.9% and 0.2% of yield variation, respectively (**Tab. 1**).

Tab. 1: Variance component estimates and percentages for yields in the OCCC barley performance trials from 1958 to 2021.

Term	Variance Component Estimate	Percentage of Variation (%)
Location × Year ($(LY)_{jk}$)	800571.696	57.0
Location (L_j)	205164.332	14.6
Residual ($(GLY)_{ijk}$)	194498.943	13.8
Year (Y_k)	148078.279	10.5
Variety (G_i)	41037.034	2.9
Variety × Year ($(GY)_{ik}$)	12200.023	0.9
Variety × Location ($(GL)_{ij}$)	3039.387	0.2

Both genetic and environmental factors accounted for increases in historical yields

The estimated rate of yield increase as explained by genetic factors was 16.8 kg ha^{-1} $year^{-1}$ ($P < 0.001$) (**Fig. 3**). The average adjusted yield means due to genetic factors was 3482.516 kg ha^{-1} from 1958 to 1989, whereas it was 4019.255 kg ha^{-1} from 1990 to 2021.

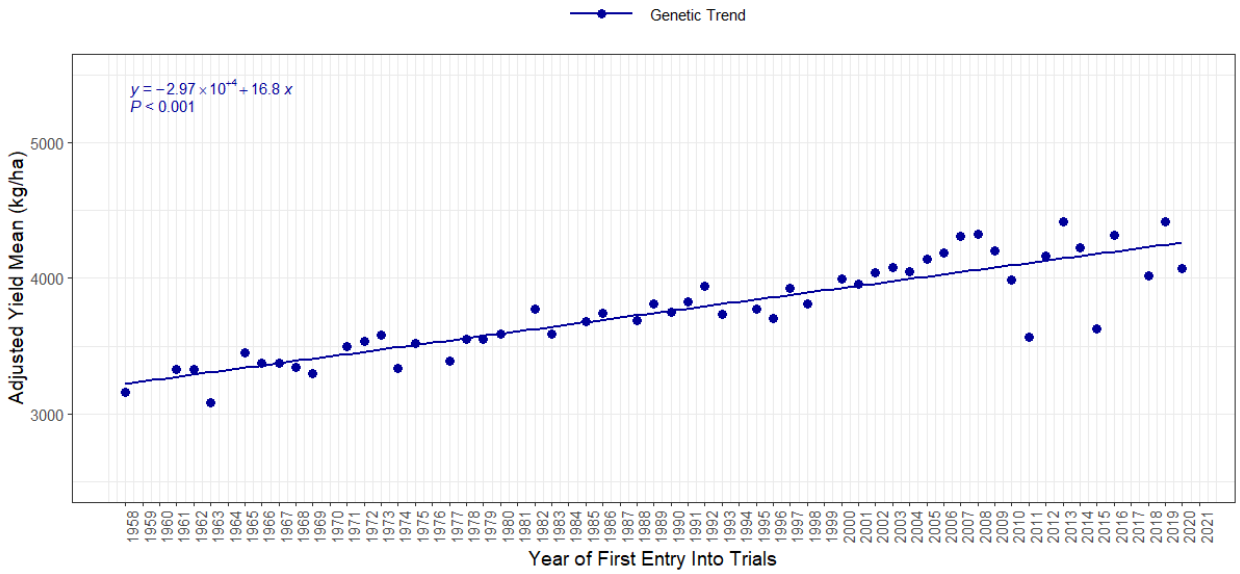

Fig. 3: Genetic trend for historical barley yields, obtained by plotting the adjusted yield means due to genetic factors against the YFE of a variety.

The estimated rate of yield increase as explained by environmental factors, which include agronomic and climatic changes, was lower at 13.6 kg ha^{-1} $year^{-1}$ ($P < 0.001$) (**Fig. 4**). The average adjusted yield means due to environmental factors was 3523.143 kg ha^{-1} from 1958 to 1989 and 4062.863 kg ha^{-1} from 1990 to 2021.

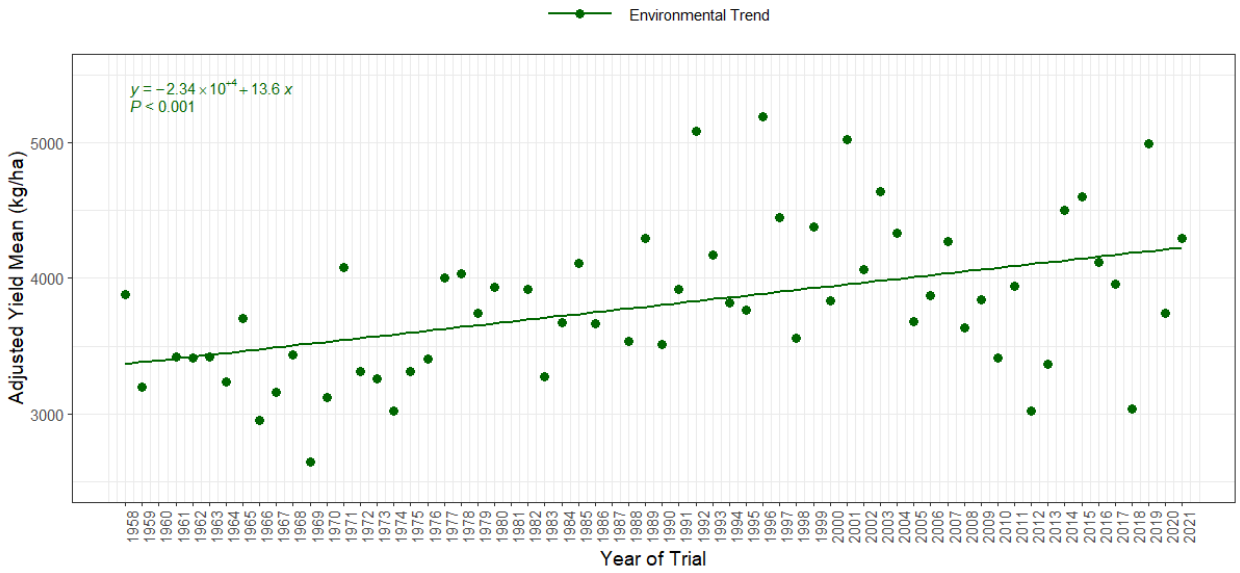

Fig. 4: Environmental trend for historical barley yields, obtained by plotting the adjusted yield means due to environmental factors against the year of trial.

Yield differences among varieties correlated with genome-wide allelic differences

Yield changes among varieties also correlated with genome-wide allelic changes. Performing PCA on 7,813 SNPs revealed that principal component 1 (PC1) explained 26.2% and PC2 explained 9.33% of the total allelic variation. Moreover, PC1 captured alleles that determine barley row type (**Fig. 5**, **Fig. 6**), two of the most prominent of which are six-row and two-row (Canadian Grain Commission, 2023). On the other hand, PC2 captured allelic differences in older and newer varieties (**Fig. 6**). Plotting PC1 values against PC2 values revealed that six-row and two-row varieties formed distinct clusters; similarly, older and newer varieties formed clusters, albeit less distinct (**Fig. 5**). Nevertheless, genome-wide allelic changes were observed over time in both six- and two-row varieties (**Fig. 5**).

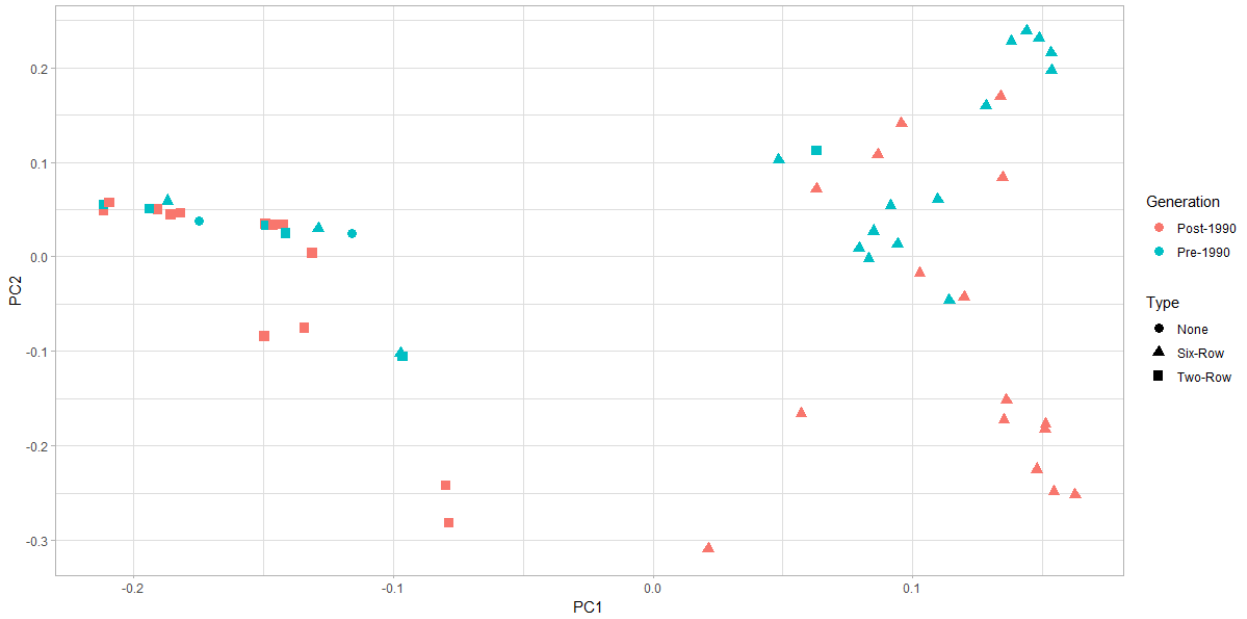

Fig. 5: Plot of PC1 values against PC2 values, where entries were differentiated based on whether they belonged to old, new, six-row, and two-row varieties.

Plotting PC1 values against varieties' adjusted yield means showed that yield improvements over time occurred in both six-row and two-row varieties (**Fig. 6**), supporting the positive trend due to genetic factors (**Fig. 3**).

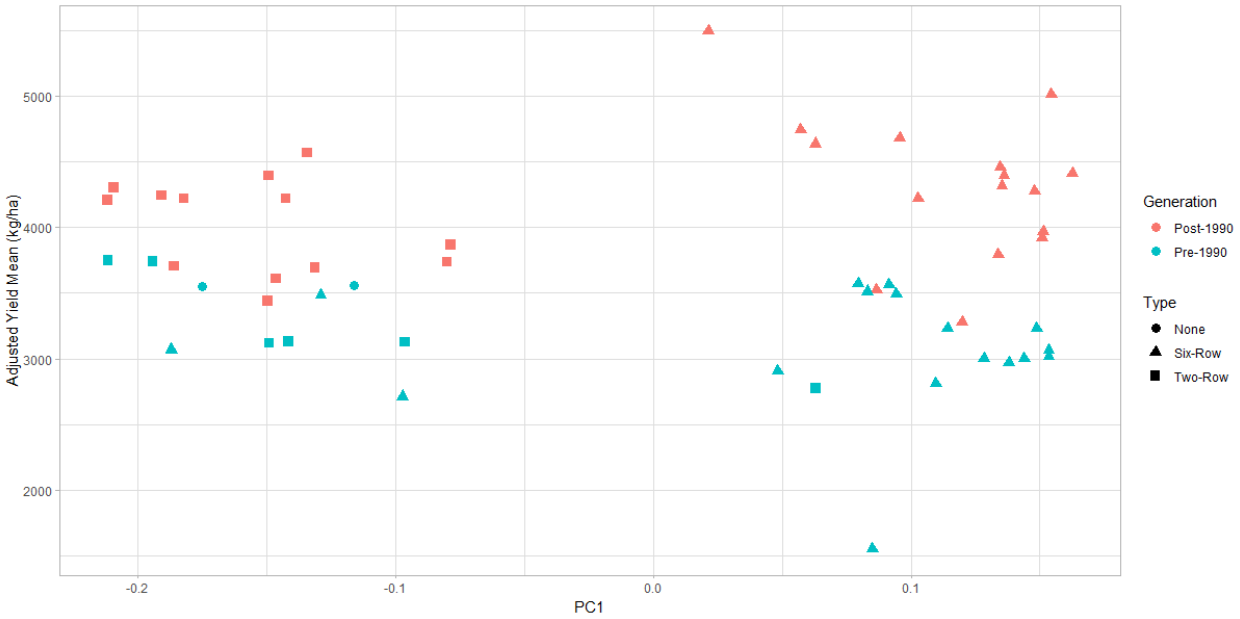

Fig. 6: Plot of PC1 values against varieties' adjusted yield means, where entries were differentiated based on whether they belonged to old, new, six-row, and two-row varieties.

Genomic prediction models predicted varieties' adjusted yield means based on their genotype with moderate accuracy

Adjusted yield means predicted from the genotypes of overlapping varieties were obtained across all 150 iterations of 5-fold cross-validation. Then, the average predicted adjusted yield means were calculated per variety. The average predicted adjusted yield means correlated moderately with the observed adjusted yield means, with a mean prediction accuracy of 0.557 (**Fig. 7**) and a standard deviation of 0.062 (**Fig. 8**).

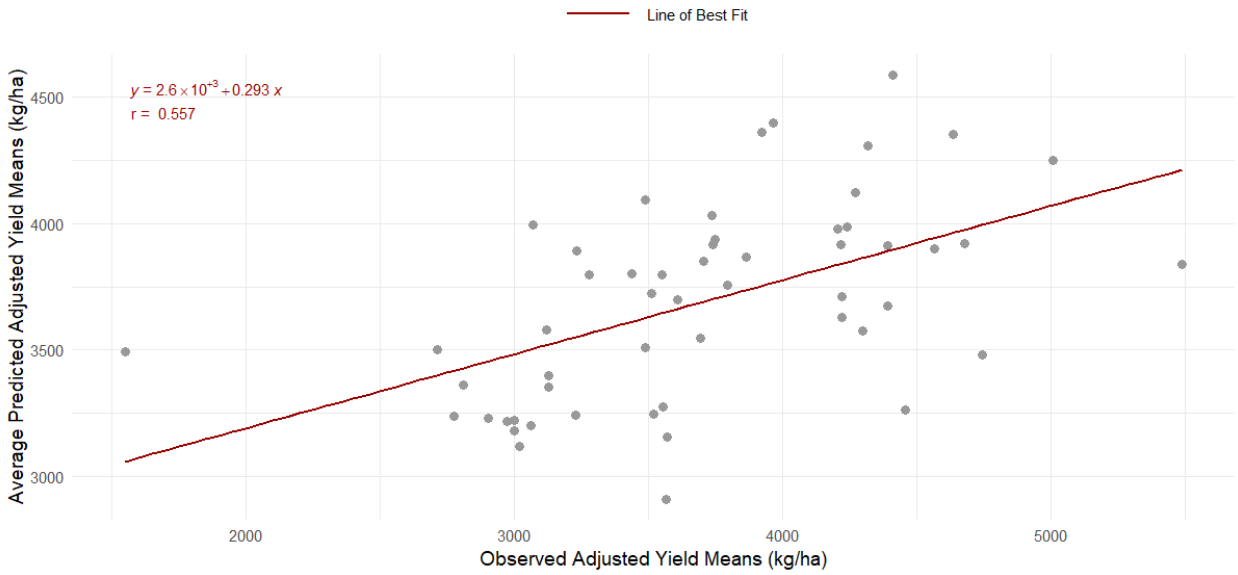

Fig. 7: Plot of the observed adjusted yield means against the average predicted adjusted yield means for the overlapping varieties.

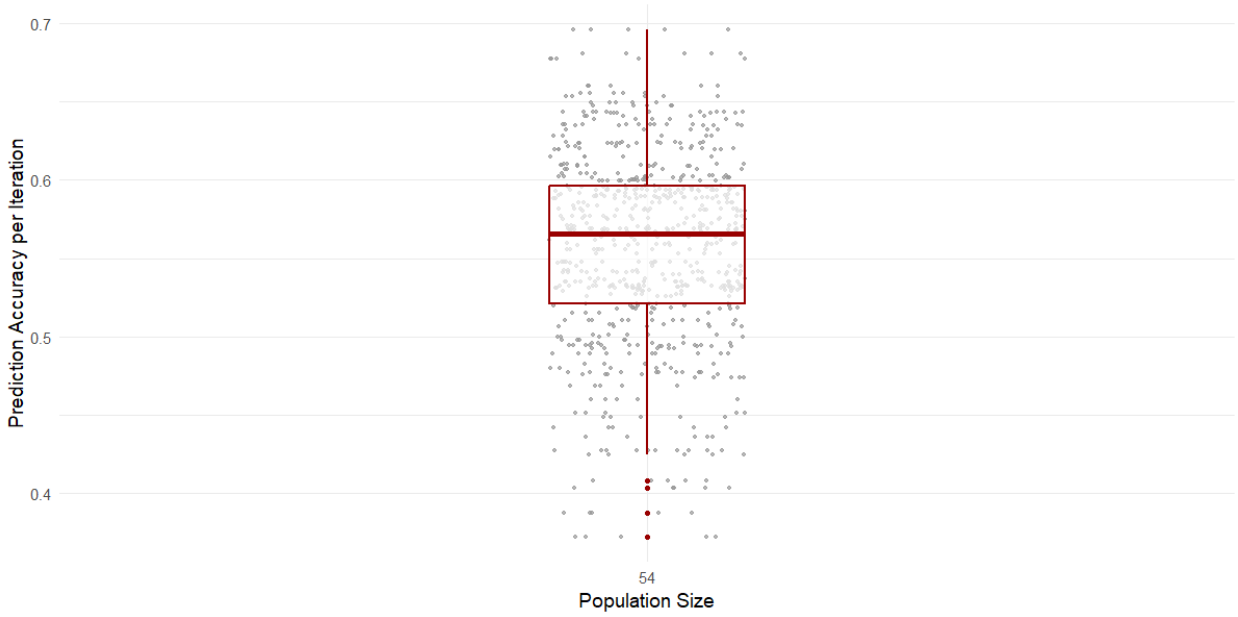

Fig. 8: Plot of the prediction accuracies of the genomic prediction model per iteration.

Old varieties' genotype information was not predictive of new varieties' yields and vice versa

Genotype information from old varieties correlated very weakly with yields of new varieties, with a mean correlation of -0.122 (**Fig. 9**). The same is true for predicting the yields of old varieties using genotype information from new varieties, as seen by the mean correlation of 0.016 (**Fig. 10**).

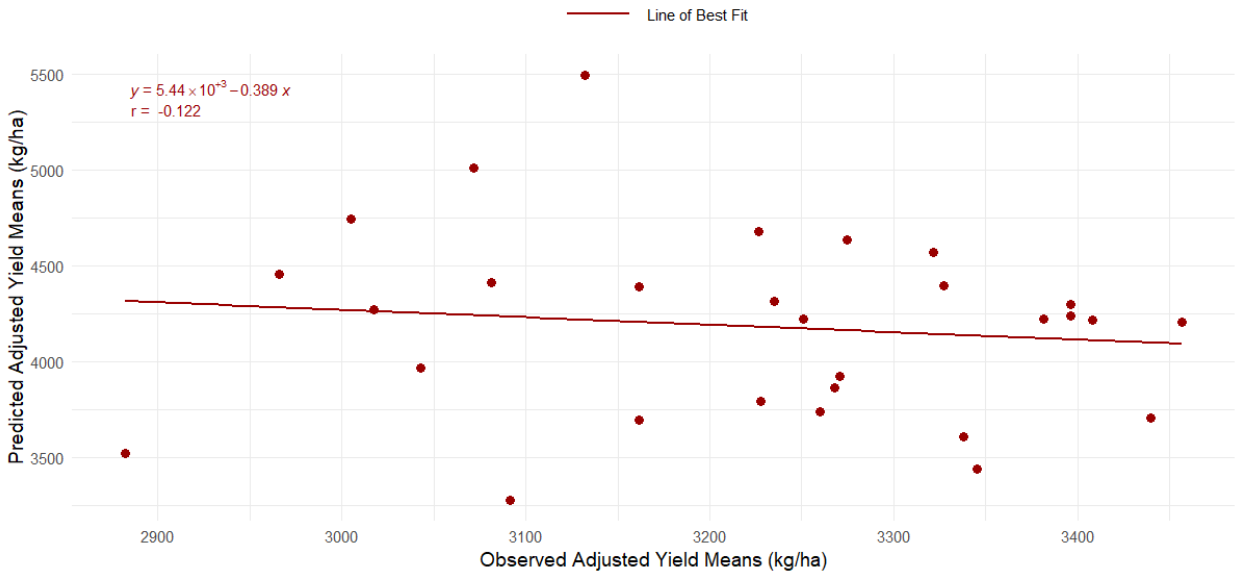

Fig. 9: Plot of the observed adjusted yield means against the predicted adjusted yield means when using genotype information from old varieties to predict yields of new varieties.

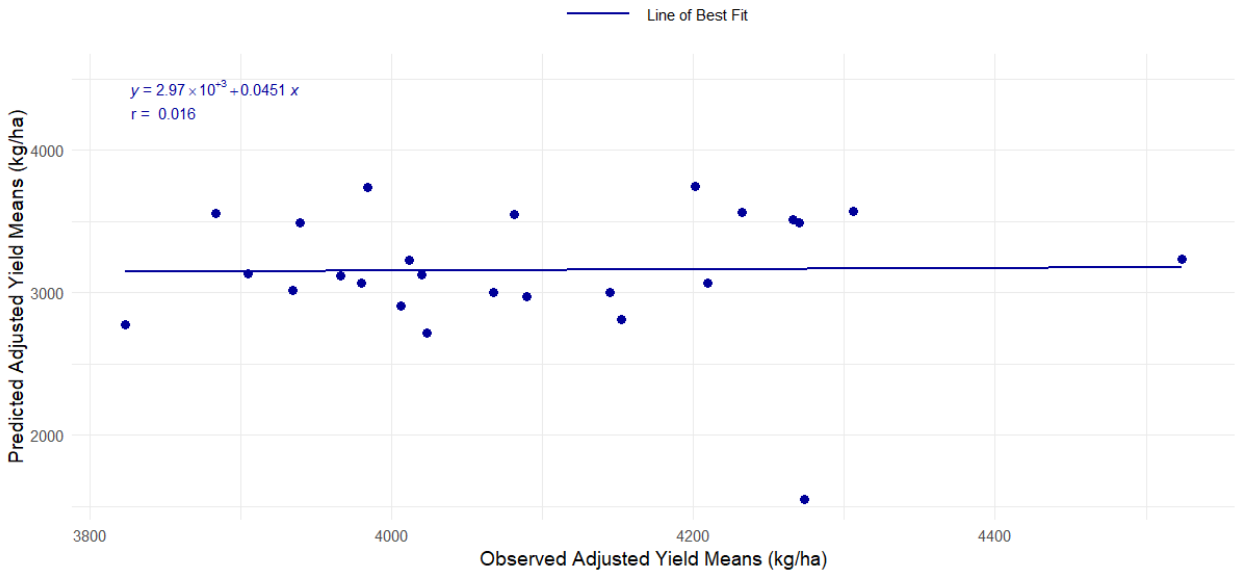

Fig. 10: Plot of the observed adjusted yield means against the predicted adjusted yield means when using genotype information from new varieties to predict yields of old varieties.

Discussion

Ontarian barley yields have steadily increased, albeit not as rapidly as in other countries

The yields of Ontarian barley varieties increased steadily from 1958 to 2021 (**Fig. 2**). However, the rate of increase is slower compared to other countries. For example, with fungicide treatment, the rates of yield increase for German six-row winter barley, two-row winter barley, and spring barley from 1983 to 2012 were 76.1 kg ha^{-1} $year^{-1}$, 76.8 kg ha^{-1} $year^{-1}$, and 48.3 kg ha^{-1} $year^{-1}$, respectively (Laidig et al., 2014). In the Netherlands, spring barley yields have increased at a rate of 70 kg ha^{-1} $year^{-1}$ from around 1980 to 2010 (Rijk et al., 2013).

Increasing fungicide use may help further expedite yield gains by combating foliar diseases (Bandara et al., 2020). So et al. (2022) reported that in Ontarian winter wheat variety trials, fungicide treatment observed within intensive trials conferred yield increases equaling 19.7 years of genetic improvement within non-intensive trials. Additionally, from looking at 436 studies investigating the effectiveness of fungicide use on foliar disease and yield, Wise et al. (2019) found that the overall yield response to fungicide use was significant ($P < 0.001$) at 332.9 ± 29.1 kg ha^{-1}. However, an important consideration for widespread fungicide use is the risk of selecting fungicide-resistant pathogen strains (Bandara et al., 2020). Future studies should therefore investigate the effects of fungicide use on Ontarian barley yields and its magnitude, as well as any considerations for and the potential risks of widespread fungicide application.

Environment accounted for yield increases but also a high amount of yield variation, while genetics accounted for mostly yield increases only

Environmental factors encompass agricultural and climatic factors that affect the performance of various locations and years. The increase in yields due to the environment was estimated to be 13.6 kg ha^{-1} $year^{-1}$, although year-to-year yields can vary significantly (**Fig. 4**). Variance component estimates further corroborate the fact that the environment accounted for most of yield variation. For instance, the location-by-year interaction effects, locations, and years explained 57%, 14.6%, and 10.5% of yield variation, respectively (**Tab. 1**). Pertinent environmental factors include delayed planting (Tapley et al., 2013), high temperatures during the summer, and high disease pressure (So et al., 2022). Nonetheless, these factors have little utility in predicting future crop performance (So et al., 2022). Similarly, the increase in yields

due to genetic factors was estimated to be 16.8 kg ha^{-1} $year^{-1}$, but it is much more consistent year to year (**Fig. 3**). This consistency is supported by the fact that varieties, the variety-by-year interaction effects, and the variety-by-location interaction effects explained a mere 2.9%, 0.9%, and 0.2% of yield variation, respectively (**Tab. 1**). Future efforts at increasing yields should therefore focus on breeding genetically superior varieties.

Yield increases over time correspond to genome-wide allelic shifts, and genomic prediction has the potential to accelerate future progress

Genome-wide allelic shifts occurred in both six-row and two-row varieties (**Fig. 5**), and these changes correspond to yield increases over time (**Fig. 6**). Because of these allelic shifts, marker estimations from old varieties could not predict the yields of new varieties (**Fig. 9**) and vice versa (**Fig. 10**). Nevertheless, genomic prediction has the potential to accelerate future progress by estimating the performance of an untested variety from its genotype (**Fig. 7**, **Fig. 8**). Future studies should, however, incorporate genotype data from a more comprehensive list of varieties to enhance the applicability of genomic prediction in increasing barley production in Ontario. It is also important to include a greater number of entry replicates to improve the accuracy of genomic prediction models (Lorenz, 2013).

References

Andrews, S. (2010, April 26). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics. **https://www.bioinformatics.babraham.ac.uk/projects/fastqc/**

Auton, A., & Marcketta, A. (2015). VCFtools. **https://vcftools.github.io/man_0112a.html**

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G., & Schroeder, J. (2019). Genetic strategies for improving crop yields. *Nature, 575*(7781), 109–118. **https://doi.org/10.1038/s41586-019-1679-0**

Bandara, A. Y., Weerasooriya, D. K., Conley, S. P., Bradley, C. A., Allen, T., & Esker, P. D. (2020). Modeling the relationship between estimated fungicide use and disease-associated yield losses of soybean in the United States I: Foliar fungicides vs foliar diseases. *PLOS ONE, 15*(6), e0234390. **https://doi.org/10.1371/journal.pone.0234390**

Bartholomé, J., Prakash, P. T., & Cobb, J. N. (2022c). Genomic Prediction: Progress and Perspectives for rice improvement. In *Methods in molecular biology* (pp. 569–617). **https://doi.org/10.1007/978-1-0716-2205-6_21**

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., & Tanaka, E. (2023, November 5). *Package 'lme4'*. The Comprehensive R Archive Network. **https://cran.r-project.org/web/packages/lme4/lme4.pdf**

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics, 23*(19), 2633–2635. **https://doi.org/10.1093/bioinformatics/btm308**

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics, 103*(3), 338–348. **https://doi.org/10.1016/j.ajhg.2018.07.015**

Canadian Grain Commission. (2023, August 1). *Barley: Classes and varieties*. **https://www.grainscanada.gc.ca/en/grain-quality/official-grain-grading-guide/06-barley/classes-types-varieties.html**

Danecek, P., Auton, A., Abecasis, G. R., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156–2158. **https://doi.org/10.1093/bioinformatics/btr330**

Endelman, J. (2023, January 6). *Package 'rrBLUP'*. The Comprehensive R Archive Network. **https://cran.r-project.org/web/packages/rrBLUP/rrBLUP.pdf**

European Nucleotide Archive. (n.d.). *Assembly: GCA_904849725.1*. **https://www.ebi.ac.uk/ena/browser/view/GCA_904849725.1**

Government of Canada. (2022, June 7). *Varieties of Crop Kinds Registered in Canada*. **https://active.inspection.gc.ca/netapp/regvar/regvar_lookupe.aspx**

Jackson, R. J., Buntjer, J. B., Bentley, A. R., Da Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., Poupard, B., Smith, S. P., Hayes, C., Barber, T., Love, B., Gaynor, R. C., Gorjanc, G., Howell, P., Mackay, I., Hickey, J. M., & Ober, E. S. (2023c). Phenomic and genomic prediction of yield on multiple locations in winter wheat. *Frontiers in Genetics, 14*. **https://doi.org/10.3389/fgene.2023.1164935**

Laidig, F., Drobek, T., & Meyer, U. (2008). Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. *Plant Breeding, 127*(6), 541-547. **https://doi.org/10.1111/j.1439-0523.2008.01564.x**

Laidig, F., Piepho, H., Drobek, T., & Meyer, U. (2014). Genetic and non-genetic long-term trends of 12 different crops in German official variety performance trials and on-farm yield trends. *Theoretical and Applied Genetics, 127*(12), 2599–2617. **https://doi.org/10.1007/s00122-014-2402-z**

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023, October 18). *Package 'emmeans'*. The Comprehensive R Archive Network. **https://cran.r-project.org/web/packages/emmeans/emmeans.pdf**

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987–2993. **https://doi.org/10.1093/bioinformatics/btr509**

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics, 26*(5), 589–595. **https://doi.org/10.1093/bioinformatics/btp698**

Lorenz, A. J. (2013). Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3: Genes, Genomes, Genetics, 3*(3), 481–491. **https://doi.org/10.1534/g3.112.004911**

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal, 17*(1), 10. **https://doi.org/10.14806/ej.17.1.200**

najoshi. (2013, September 27). *sabre - A barcode demultiplexing and trimming tool for FastQ files*. GitHub. **https://github.com/najoshi/sabre**

Ontario Ministry of Agriculture, Food and Rural Affairs. (2023, October 23). *Cereal Production in Ontario*. **https://omafra.gov.on.ca/english/crops/field/cereal.html**

Piepho, H., Laidig, F., Drobek, T., & Meyer, U. (2014). Dissecting genetic and non-genetic sources of long-term yield trend in German official variety trials. *Theoretical and Applied Genetics, 127*(5), 1009–1018. **https://doi.org/10.1007/s00122-014-2275-1**

Poland, J., Brown, P. J., Sorrells, M. E., & Jannink, J. (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. **https://doi.org/10.1371/journal.pone.0032253**

Rijk, B., Van Ittersum, M., & Withagen, J. (2013). Genetic progress in Dutch crop yields. *Field Crops Research, 149*, 262–268. **https://doi.org/10.1016/j.fcr.2013.05.008**

Rimmer, A. (2017). Platypus 0.1.5 documentation. **https://rahmanteamdevelopment.github.io/Platypus/documentation.html**

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A., McVean, G., & Gerton, L. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics, 46*(8), 912–918. **https://doi.org/10.1038/ng.3036**

Smith, A. (2023, June 21). *Rough guide for merging vcfs* [Word document]. Microsoft Corporation.

So, D. (n.d.). *Chapter 3 - Trend Modeling* [R script]. The R Foundation.

So, D., Smith, A., Sparry, E., & Lukens, L. (2022). Genetics, not environment, contributed to winter wheat yield gains in Ontario, Canada. *Theoretical and Applied Genetics, 135*(6), 1893–1908. **https://doi.org/10.1007/s00122-022-04082-3**

Strandén, I., & Christensen, O. (2011). Allele coding in genomic evaluation. *Genetics Selection Evolution, 43*(1). **https://doi.org/10.1186/1297-9686-43-25**

Tapley, M., Ortiz, B. V., Van Santen, E., Balkcom, K. S., Mask, P. L., & Weaver, D. (2013). Location, seeding date, and variety interactions on winter wheat yield in southeastern United States. *Agronomy Journal, 105*(2), 509–518. **https://doi.org/10.2134/agronj2012.0379**

Tougui, I., Jilbab, A., & Mhamdi, J. E. (2021). Impact of the choice of Cross-Validation techniques on the results of Machine Learning-Based diagnostic Applications. *Healthcare Informatics Research, 27*(3), 189–199. **https://doi.org/10.4258/hir.2021.27.3.189**

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., & Belzile, F. (2017). Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics, 18*(1). **https://doi.org/10.1186/s12859-016-1431-9**

Wise, K., Smith, D. L., Freije, A., Mueller, D. S., Kandel, Y. R., Allen, T., Bradley, C. A., Byamukama, E., Chilvers, M. I., Faske, T., Friskop, A., Hollier, C. A., Jackson-Ziems, T. A., Kelly, H., Kemerait, B., Price, P. P., Robertson, A. E., & Tenuta, A. (2019). Meta-analysis of yield response of foliar fungicide-treated hybrid corn in the United States and Ontario, Canada. *PLOS ONE, 14*(6), e0217510. **https://doi.org/10.1371/journal.pone.0217510**

Yang, F., Liu, Z., Wang, Y., Wang, X., Zhang, Q., Han, Y., Zhao, X., Pan, S., Yang, S. X., Wang, S., Zhang, Q., Qiu, J., & Wang, K. (2023). A variety test platform for the standardization and data quality improvement of crop variety tests. *Frontiers in Plant Science, 14*. **https://doi.org/10.3389/fpls.2023.1077196**

Zhao, H., Lin, Z., Khansefid, M., Tibbits, J., & Hayden, M. J. (2023). Genomic prediction and selection response for grain yield in safflower. *Frontiers in Genetics, 14*. **https://doi.org/10.3389/fgene.2023.1129433**

Zheng, X. (2018, March 20). *Tutorials for the R/Bioconductor Package SNPRelate*. Bioconductor. **https://www.bioconductor.org/packages/release/bioc/vignettes/SNPRelate/inst/doc/SNPRelate.html**

Additional File(s)

Commented code and project workflow

All the code used in this project, which includes shell and R scripts, as well as the documentation for the project's workflow can be found in a GitHub repository at https://github.com/ajauwena/binf_6999_project.