Preamble

Wheat (Triticum aestivum L.), responsible for a significant portion of global caloric intake, is increasingly threatened by both natural and anthropogenic factors, including climate change and geopolitical unrest (Zhang et al. 2022). These challenges underline the urgency for innovative approaches in wheat breeding, beyond the capabilities of traditional methods. Genomic Selection (GS) represents a transformative approach, leveraging high-throughput genotyping technologies to overcome the limitations of conventional breeding by enabling more precise genetic selection for desired traits(Heffner et al. 2009).

Several factors can influence the prediction accuracy of GS models, including population structure, training population size, and the choice of prediction models (Plavšin et al. 2021). Selecting an appropriate prediction model is a relatively straightforward approach to improve prediction accuracy among these factors. Traditional GS models like Genomic Best Linear Unbiased Prediction (GBLUP) and Ridge Regression (RR) have been foundational, yet they often fail to capture non-linear genetic interactions critical for complex traits like yield (Desta and Ortiz 2014). Emerging models such as Random Forest (RF) and Reproducing Kernel Hilbert Space (RKHS) offer potential improvements by accounting for these complexities . This study aims to explore the predictive capabilities of these advanced models, focusing exclusively on wheat yield (YLD) as a primary trait of interest.

The data for this analysis stems from a comprehensive dataset encompassing 250 spring wheat lines from diverse genetic backgrounds, evaluated under varying environmental conditions. Six statistical models, namely Ridge Regression (RR), Reproducing Kernel Hilbert Space (RKHS), Genomic Best Linear Unbiased Prediction (GBLUP), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), and Random Forest (RF), were evaluated to identify the optimal model for YLD. Through rigorous analysis and visualization of the genetic predictors of yield, we anticipate fostering a deeper understanding of GS's potential to address pressing food security challenges.

Data Source Description

Plant Material

This study involves 250 spring wheat varieties and elite lines, developed by breeding programs in the Northwestern Pacific region of the United States and the International Maize and Wheat Improvement Center (CIMMYT) in Mexico City, Mexico. The collection includes three market categories of spring wheat cultivated in the Americas: soft white spring, hard white spring, and hard red spring, with over 50% of the lines serving as foundational lines for regional variety development programs.

Phenotypic Evaluation

The wheat panel was evaluated across five distinct environments, labeled E1 to E5 (E1 = 2016, E2 = 2017, E3 = 2021, E4 = 2022, E5 = 2023), at a field trial location in Aberdeen, Idaho (42°56'36" N and 112°50'22" W). The trial layout was a randomized complete block design with two replicates. Each genotype was planted in 3.0 m plots consisting of seven rows, with a row spacing of 21 cm. Yield per hectare (kg/ha) was measured in four of these trials (E1, E2, E4, E5).

Charmet et al.(2020)

## Genotyping and Structure

The population was genotyped using Illumina's 90K iSelect SNPchip. Raw data was provided by the USDA/ARS Cereal Crops Research Unit and was analyzed using Genome Studio v2.0.5(Illumina 2010) for processing its raw data. Polymorphic markers were identified as exhibiting clear, distinct clusters in Genome Studio, with a minimum distance of 0.20 between the polar coordinates of normalized theta intensities. Markers were further filtered based on missing data (>10%) and minor allele frequency (<5%) using TASSEL v5.2.89 (Bradbury et al. 2007).

Data Dictionary

Results

Yield (YLD) is a critical agronomic trait determining the superiority of wheat varieties. The BLUPs for YLD ranged from 4574.49 to 6827.75 kg/ha. Lowest YLD was observed in E2, with a mean YLD of 5528.63 kg/ha, and highest YLD was observed in E5, with a mean YLD of 6604.10 kg/ha. YLD in E5 was on average 1075.47 kg/ha higher than in E2. The distribution YLD demonstrated the most significant variability, with a left-skewed and leptokurtic distribution. The variability of YLD data varied significantly across environments, indicating that yield is greatly influenced by environmental factors (Fig. 1)

The BLUPs values for YLD trait showed high correlations with the trait values across different environments, indicating their effectiveness in integrating phenotypic data from multiple environments for subsequent genetic analysis. Although the correlations between YLD BLUPs and YLD in some environments were slightly lower, they still represented the overall performance of YLD reasonably well (Fig. 2)

**Population Structure**

Principal component analysis (PCA) based on the genotypic data revealed a clear population structure among the 249 spring wheat lines (Fig. 5). The first three principal components (PCs) explained 6.3%, 5.4%, and 4.4% of the total genetic variance, respectively. The 3D PCA plot showed distinct clustering of lines according to their sources. The number of lines within each cluster ranged from 10 (Other) to 111 (UI). The largest cluster consisted of 111 lines from the University of Idaho (UI), followed by 34 lines from UC Davis (UCD) and 26 lines from Washington State University (WSU). Analysis of variance (ANOVA) revealed that the population structure had significant effects on all five traits analyzed (P < 0.001) (Supplemental Table 4). Among these traits, thousand kernel weight (TKW) showed the highest proportion of variance explained by the clusters (R2 = 23.6%), followed by plant height (PHT, R2 = 21.0%), grain yield (YLD, R2 = 13.2%), total spikelet number per spike (tSNS, R2 = 9.8%), and heading date (HD, R2 = 7.4%).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Df | Sum Sq | Mean Sq | F value | Pr(>F) | R\_Squared | Trait |
| 6 | 1383462.467 | 230577.0778 | 5.068379905 | 6.41E-05 | 0.111634111 | YLDall |
| 242 | 11009366.68 | 45493.25073 | NA | NA | NA | YLDall |

**Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, and Buckler ES**. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007:**23**(19):2633–2635. https://doi.org/10.1093/bioinformatics/btm308

**Charmet G, Tran L-G, Auzanneau J, Rincent R, and Bouchet S**. BWGS: AR package for genomic selection and its application to a wheat breeding programme. PloS one. 2020:**15**(4):e0222733. https://doi.org/10.1371/journal.pone.0222733

**Desta ZA and Ortiz R**. Genomic selection: genome-wide prediction in plant improvement. Trends in plant science. 2014:**19**(9):592–601. https://doi.org/10.1016/j.tplants.2014.05.006

**Heffner EL, Sorrells ME, and Jannink J-L**. Genomic Selection for Crop Improvement. Crop Science. 2009:**49**(1):1–12. https://doi.org/10.2135/cropsci2008.08.0512

**Illumina**. Introduction to GenomeStudio Software. 2010.

**Plavšin I, Gunjača J, Šatović Z, Šarčević H, Ivić M, Dvojković K, and Novoselović D**. An overview of key factors affecting genomic selection for wheat quality traits. Plants. 2021:**10**(4):745. https://doi.org/10.3390/plants10040745

**Zhang T, He Y, DePauw R, Jin Z, Garvin D, Yue X, Anderson W, Li T, Dong X, Zhang T, et al.** Climate change may outpace current wheat breeding yield improvements in North America. Nature Communications. 2022:**13**(1):5591. https://doi.org/10.1038/s41467-022-33265-1