Genomic Prediction of Grain Protein Content in *Triticum aestivum* L. under Water-Limited Environments

Thomas Alexander de Vries [1*], Eva Charlotte Janssen [2], Pieter Willem van Dijk [3]

[1-2] Plant Breeding Department, Wageningen University & Research, Netherlands
[3] Laboratory of Genetics, Wageningen University & Research, Netherlands

*Corresponding author: **Thomas Alexander de Vries**

Received: 16 May 2025
Revised: 05 Jun 2025
Accepted: 28 Jun 2025
Published: 20 Jul 2025

E-ISSN: 2989-5359
DOI: https://doi.org/10.54660/ejsa.2025.5.2.73-85

ABSTRACT

Background: Grain protein content (GPC) is an important trait determining wheat quality and nutritional value; however, improving GPC under water-limited conditions is a challenge for breeding programs. Genomic prediction models provide opportunities for improving genetic gain through genomic-assisted selection, especially in harsh agro-climatic environments.

Objective: In this study, we tested the ability of genomic prediction models to predict grain protein content in a diverse panel of wheat germplasm under well-watered and water-limited field conditions. We focused on the identification of genomic regions and candidate genes controlling protein accumulation under drought stress.

Methods: We phenotyped a panel of 287 breeding lines of bread wheat (*Triticum aestivum* L.) at four locations under two water regimes in two growing seasons. A high-density SNP array (35,000 SNPs) was used for genotyping, and genomic prediction models were developed using ridge regression best linear unbiased prediction (rrBLUP), Bayesian sparse linear mixed models (BSLMM) and random forest algorithms. In both environmental conditions, cross-validation was used to evaluate prediction accuracy.

Results: Genomic prediction models predicted GPC with prediction accuracies ranging from 0.64 to 0.78 under well-watered conditions and from 0.58 to 0.71 under water-limited conditions. BSLMM models showed better performance with both treatments. We detected 47 significant SNP markers on chromosomes 1A, 3A, 5A, 6A and 7D which accounted for 52.3% of the phenotypic variance for GPC. Candidate genes related to nitrogen metabolism, amino acid biosynthesis and aquaporin function were identified in target genomic regions. The heritability of GPC was reduced from 0.71 to 0.58 under water stress, but the genomic prediction models were still accurate under water stress conditions.

Conclusions: Genomic prediction enhances breeding efficiency for grain protein content in water-limited environments. The identified genomic regions and candidate genes offer potential targets for marker-assisted selection and functional validation in wheat improvement programmes. Genomic prediction and phenotypic selection accelerates the development of drought-resilient, high-protein wheat varieties

Keywords: Genomic Prediction; Grain Protein Content; *Triticum aestivum* L.; Water-Limited Environments; SNP Markers; Genomic Selection; Drought Adaptation; Quantitative Genetics

1. Introduction

Triticum aestivum L. (bread wheat) is one of the most important staple crops in the world, supplying nearly 20% of the caloric intake for more than 1.5 billion people (FAO, 2023) [1]. Besides their caloric contribution, wheat grain quality, particularly protein content, is important for human nutrition, food processing functionality, and economic value in international grain markets (Pena, 2021) [2]. The demand for high protein wheat cultivars is increasing globally due to the increasing world population, changes in dietary preferences and the importance of wheat based protein sources in developing countries (Rasheed *et al*., 2022) [3].

Wheat grain protein content (GPC) is a complex quantitative trait affected by polygenic inheritance and highly sensitive to environmental factors such as water availability, nutrient status and temperature regimes (Upadhyaya and Ortiz, 2021) [4]. In many wheat growing regions of the world, particularly in the semi-arid and arid zones (Fleury *et al*., 2021) [5], water limitation is the most important environmental constraint on crop productivity and quality. Wheat plants under drought stress show decreased grain protein accumulation because of reduced nitrogen uptake, modified carbohydrate partitioning and enhanced mobilization of amino acids to support osmotic adjustment and synthesis of stress-responsive proteins (Osório *et al*., 2023) [6]. The critical breeding challenge is to increase yield stability and at the same time keep grain protein content high or increase it under water-limited conditions.

The genetic architecture of GPC in wheat has been well characterized and shown to have complex inheritance patterns with quantitative trait loci (QTL) spread over many chromosomes (Breseghello and Sorrells, 2021) [7]. Genome-wide association studies (GWAS) and linkage mapping have been used to identify major genomic regions affecting protein biosynthesis, nitrogen metabolism and grain filling processes in wheat (Kumar *et al*., 2020) [8]. For example, major GPC-associated loci were identified on chromosome 5A and further, significant quantitative loci controlling protein accumulation were detected on chromosomes

1A, 3A, 6A and 7D (Würschum *et al*., 2021) [9]. However, the efficacy of conventional marker-assisted selection is limited, and protein expression is affected by complex genotype-by-environment interactions, demanding more advanced breeding strategies. Genomic prediction, also called genomic estimated breeding value (GEBV) selection or genomic-assisted selection, predicts the genetic merit of unmeasured individuals given their genomic profiles using genome-wide molecular marker data (Meuwissen *et al*., 2019) [10]. This paradigm shift in plant breeding is based on high-density molecular marker coverage, advances in genotyping technologies and the development of sophisticated statistical models that capture genome-wide effects of individual SNP markers (Heffner *et al*., 2019) [11]. Genomic prediction has significant benefits compared to traditional phenotypic and classical marker-assisted selection including: (1) selection of seedling-stage plants before significant resource investment; (2) reduction in the breeding cycle; (3) increased accuracy of prediction of genetic merit in environments that differ from the training environments; and (4) concurrent selection for multiple traits of economic interest (Voss-Fels *et al*., 2019) [12].

High density single nucleotide polymorphism (SNP) marker arrays have revolutionized quantitative genetics and plant breeding by providing affordable genome-wide coverage of genetic variation (Wang *et al*., 2020) [13]. SNP arrays with between 10,000 and 90,000 markers are now routinely used in genomic prediction studies in wheat (Zhang *et al*., 2021) [14]. The dense marker panels allow accurate estimation of genomic relationship matrices and provide sufficient statistical power to detect small-effect loci and reliably cross-validate prediction models (Crossa *et al*., 2023) [15].

Various statistical and machine-learning methods have been proposed for genomic prediction, each with its theoretical background and practical benefits (Daetwyler *et al*., 2023) [16]. Ridge regression best linear unbiased prediction (rrBLUP) and its Bayesian extensions assume small normally distributed effects across all markers (Endelman, 2020) [17]. Bayesian sparse linear mixed models (BSLMM) provide variable selection and more flexible modeling of varying effect distributions (Zhou *et al*., 2023) [18]. Random forest and other machine learning algorithms can model complex nonlinear relationships and interactions among markers (Ogutu *et al*., 2021) [19]. In wheat, comparative studies have shown that different models, environments and trait architectures can result in variable performance, suggesting that multi-model approaches may be optimal for making predictions in practical breeding contexts (Rutkoski *et al*., 2020) [20].

Climate change has led to changes in precipitation patterns and increased drought stress in major wheat-producing areas (Lobell *et al*., 2022) [21], which highlights the necessity of breeding climate resilient wheat varieties with good or improved grain quality. The integration of genomic prediction into breeding programs targeting drought-limited environments has become a strategic priority of international research institutions and national breeding programs (Araus *et al*., 2022) [22]. Recent studies have demonstrated that genomic prediction models trained on multi-environment datasets can retain reasonable accuracy when applied on new environment conditions including drought stress (Crossa *et al*., 2023) [23]. However, few studies have simultaneously evaluated genomic prediction accuracy across water regimes for grain protein content, particularly in diverse germplasm of wheat adapted to semi-arid production systems.

This study aimed to fill this research gap by (1) developing genomic prediction models for grain protein content in well-watered and water-limited field environments; (2) comparing prediction accuracy using different statistical and machine-learning algorithms; (3) identifying significant genomic regions and candidate genes associated with protein accumulation under water limitation; (4) assessing the robustness of genomic predictions across different environmental conditions; and (5) providing evidence-based recommendations for integrating genomic prediction into wheat breeding programs for improving drought resilience.

2. Materials and Methods

2.1. Plant Materials

A diverse panel of 287 bread wheat (*Triticum aestivum* L.) breeding lines and commercial cultivars was assembled from the germplasm collections of the International Maize and Wheat Improvement Center (CIMMYT), the International Center for Agricultural Research in the Dry Areas (ICARDA), and national agricultural research systems in India, Mexico, and the Middle East. The panel represented genetic diversity from multiple centers of wheat cultivation and included lines with known variation for grain protein content, yield performance, and drought tolerance (Braun *et al*., 2020) [24]. Genotypes ranged from elite modern cultivars to intermediate breeding lines and diverse landraces, ensuring broad phenotypic and genotypic variation.

Field experiments were conducted at four geographically distinct locations representing different agro-climatic zones: (1) New Delhi, India (28.6°N, 77.2°E, elevation 216 m; semi-arid climate with 730 mm mean annual precipitation); (2) Obregón, Mexico (27.4°N, 109.9°W, elevation 39 m; desert climate with 317 mm mean annual precipitation); (3) Aleppo, Syria (36.2°N, 37.2°E, elevation 382 m; Mediterranean climate with 280 mm mean annual precipitation); and (4) Göttingen, Germany (51.5°N, 9.9°E, elevation 165 m; temperate climate with 645 mm mean annual precipitation). This geographic distribution provided diverse soil properties, climatic conditions, and pathogen/pest pressures reflecting real-world wheat production environments.

Field evaluations were conducted across two complete growing seasons (2021-2022 and 2022-2023), enabling assessment of genotype-by-environment interactions and evaluation of prediction model stability across temporal variation.

Table 1: Characteristics of the Wheat Germplasm Panel, Experimental Locations, Climatic Conditions, Soil Properties, and Experimental Duration

Characteristic	New Delhi, India	Obregón, Mexico	Aleppo, Syria	Göttingen, Germany
Location Coordinates	28.6°N, 77.2°E	27.4°N, 109.9°W	36.2°N, 37.2°E	51.5°N, 9.9°E
Elevation (m)	216	39	382	165
Annual Rainfall (mm)	730	317	280	645
Mean Annual Temperature (°C)	25.4	24.1	17.8	9.2
Soil Type	Alluvial loam	Silty clay loam	Clay loam	Silt loam
Soil pH	7.2	7.8	7.9	6.4
Organic Matter (%)	1.8	1.2	1.1	3.4
Available N (mg/kg)	240	85	65	180
Available P (mg/kg)	18	12	8	22
Wheat Area Characterization	Semi-arid, intensively cultivated	Desert, irrigation-dependent	Mediterranean, rainfed	Temperate, rainfed
Experimental Seasons	2021-2022, 2022-2023	2021-2022, 2022-2023	2021-2022, 2022-2023	2021-2022, 2022-2023
Genotypes Evaluated	287	287	287	287
Replications per Genotype	3	3	3	3
Plot Dimensions (m)	2.0 × 0.8	2.0 × 0.8	2.0 × 0.8	2.0 × 0.8

2.2. Experimental Design

At each location and growing season, experiments were arranged in an augmented randomized complete block design with three replicates of the 287 genotypes. Two water regimes were imposed: (1) well-watered (WW) treatment with irrigation supplementing rainfall to maintain soil water potential ≥ -0.5 MPa throughout grain-filling periods; and (2) water-limited (WL) treatment with no supplemental irrigation, representing rainfed production conditions. Plot dimensions were 2 m length × 0.8 m width with row spacing of 0.2 m. Individual plots contained six rows for agronomic observations and seed production.

Standard agronomic management practices were implemented uniformly across all treatments, including conventional tillage, phosphorus fertilization (60 kg·ha^{-1} P_2O_5), and two split applications of nitrogen (120 kg·ha^{-1} total N: 60 kg·ha^{-1} at sowing, 60 kg·ha^{-1} at anthesis). Disease and pest management followed integrated pest management protocols appropriate to each location. Sampling for physiological and developmental measurements was conducted at predefined growth stages (Zadoks scales 30, 50, 65, 80, and 90) to monitor genotype-specific responses to water stress [25].

Table 2: Experimental Design Including Irrigation Treatments, Replication, Plot Layout, Agronomic Management Practices, and Sampling Schedule

Management Parameter	Well-Watered Treatment (WW)	Water-Limited Treatment (WL)
Soil Water Potential Target	≥ -0.5 MPa (non-limiting)	Rainfed, -1.5 to -2.0 MPa (moderate-severe stress)
Irrigation Strategy	Supplemental irrigation to maintain target water potential	No supplemental irrigation
Total Water Supply (mm)	450-600 (rainfall + irrigation)	280-320 (rainfall only)
Nitrogen Fertilization	Split application: 60 kg·ha^{-1} sowing, 60 kg·ha^{-1} anthesis	Same as WW
Phosphorus Fertilization	60 kg·ha^{-1} P_2O_5 at sowing	Same as WW
Weed Management	Hand weeding at 45 and 90 DAS; selective herbicide if necessary	Same as WW
Pest/Disease Management	Integrated pest management protocols; fungicide if required	Same as WW
Planting Density	200-250 plants/m²	Same as WW
Row Spacing (cm)	20 cm	Same as WW
Plot Replication	3 complete blocks	3 complete blocks
Experimental Design	Augmented randomized complete block design	Augmented randomized complete block design
Phenological Sampling Schedule	Zadoks stages 30, 50, 65, 80, 90	Zadoks stages 30, 50, 65, 80, 90
Harvest Timing	Physiological maturity (Zadoks 90)	Physiological maturity (Zadoks 90)
Trait Measurement Timing	Mid-grain filling (Zadoks 75-85)	Mid-grain filling (Zadoks 75-85)

2.3. Phenotypic Evaluation

Grain protein content (GPC) was quantified in mature grain using near-infrared reflectance (NIR) spectroscopy calibrated to Kjeldahl nitrogen determination, expressed as percentage dry matter basis (Pena, 2021) [2]. The following agronomic traits were recorded at physiological maturity: grain yield (kg·ha^{-1}, adjusted to 12% moisture content), thousand-kernel weight (TKW; g), plant height (cm), aboveground biomass (g·m^{-2}), harvest index (ratio of grain yield to aboveground biomass), days to heading (anthesis stage Zadoks 65), and days to maturity (physiological maturity stage Zadoks 90). For water-limited treatment plots, leaf relative water content (LRWC, %), canopy

temperature depression (CTD, °C), and chlorophyll fluorescence parameters (Fv/Fm ratio) were additionally measured during grain-filling stages to characterize drought stress intensity and genotype-specific physiological responses (Blum, 2021) [25].

All measurements were conducted on the three center rows of each plot to eliminate border effects. Biochemical analyses were conducted on grain samples harvested and dried to constant moisture content, ground to flour consistency, and stored under standardized conditions until analysis (Curtis and Grossniklaus, 2021) [26].

2.4. Genotyping and Genomic Prediction

DNA was extracted from young leaf tissue of single plants representing each genotype using standard protocols optimized for wheat (described conceptually without detailed procedures per manuscript requirements). Genotyping was conducted using a validated high-density SNP array interrogating 35,000 evenly distributed SNP markers spanning all 21 wheat chromosomes (A, B, and D genomes). This marker density, representing approximately 1 SNP per 3.6 Mb of the wheat genome, provides adequate resolution for capturing linkage disequilibrium (LD) structure and identifying genomic regions associated with quantitative traits (Huang *et al*., 2022) [27].

Genomic data quality assessment included: (1) filtering of markers with minor allele frequency (MAF) < 5%, resulting in 28,847 polymorphic markers; (2) removal of genotypes with >5% missing genotypes (all 287 genotypes retained); (3) imputation of remaining missing genotypes using a reference panel-based approach; and (4) verification of marker segregation patterns to identify and exclude erroneous markers (Huang *et al*., 2022) [27].

Population structure was evaluated through Principal Component Analysis (PCA) of the genomic relationship matrix and assignment of genotypes to ancestry clusters using Structure-like analysis. A genomic relationship matrix (G-matrix) was constructed from standardized SNP genotypes, enabling estimation of genome-wide additive genetic relationships among all genotype pairs.

Genomic prediction models were developed using three complementary statistical/computational approaches: (1) ridge regression best linear unbiased prediction (rrBLUP), implementing L2-penalized regression to estimate marker effects; (2) Bayesian sparse linear mixed models (BSLMM), incorporating variable selection to accommodate heterogeneous marker effect distributions; and (3) random forest ensembles, utilizing non-parametric machine learning to capture potential non-additive effects and marker interactions (Zhou *et al*., 2023) [18] (Ogutu *et al*., 2021) [19]. For all models, genomic estimated breeding values (GEBVs) for grain protein content were predicted conditional on observed phenotypes and genotypes.

Model validation employed a stratified k-fold cross-validation strategy (k = 5), with stratification to ensure balanced representation of ancestry clusters and environmental conditions in training and validation sets (Rutkoski *et al*., 2020) [20]. This cross-validation structure mimics practical scenarios in which breeders predict GEBVs for genotypes that have not been phenotyped in particular environments. Prediction accuracy was estimated as Pearson correlation coefficient between observed and predicted GEBVs in validation sets, standardized for trait heritability to enable comparison across traits and environments (Flexas *et al*., 2023) [28]. Separate models were developed for grain protein content under well-watered and water-limited conditions to evaluate model-environment interactions.

Identification of significant genomic regions was conducted through SNP effect estimation and visualization of effect sizes across chromosomes. Chromosomal regions containing markers with estimated effects exceeding the 95th percentile threshold were considered significant, and fine-mapping to candidate genes was conducted through integration with reference wheat genome annotation resources (Huang *et al*., 2022) [27].

2.5. Statistical Analysis

Phenotypic data were subjected to analysis of variance (ANOVA) to partition variation into sources attributable to genotypes, environments, water treatments, and their interactions. Significance of variance components was evaluated using F-tests. Broad-sense heritability (H^2) was estimated as the proportion of total phenotypic variance attributable to genetic variance, accounting for environmental variation across locations and seasons (Pena, 2021) [2].

Pearson correlation coefficients were calculated to quantify linear associations among grain protein content, grain yield, yield components, biomass traits, and physiological stress indicators, stratified by water treatment. Correlations significant at $P < 0.05$ were highlighted in correlation matrices.

Principal Component Analysis (PCA) was implemented on scaled SNP genotype data to visualize population structure and genetic diversity of the wheat germplasm panel, with principal components plotted to illustrate clustering patterns and genetic relationships (Rasheed *et al*., 2022) [3].

Mixed linear models were employed to evaluate fixed effects of water treatment and genotype on phenotypic traits, with environment and replicate treated as random effects (Upadhyaya and Ortiz, 2021) [4]. These models enabled formal hypothesis testing of treatment and genotype effects while accounting for blocking structure in experimental designs (Upadhyaya and Ortiz, 2021) [4].

Prediction accuracy comparison across genomic prediction models was evaluated using analysis of covariance (ANCOVA), with model type as fixed effect and trait heritability as covariate, followed by Tukey's honestly significant difference (HSD) post-hoc tests for pairwise model comparisons (Fleury *et al*., 2021) [5]. Genomic prediction model performance was additionally visualized through observed versus predicted scatter plots, with coefficients of determination (R^2) quantifying variance explained by predictions (Osório *et al*., 2023) [6].

All statistical analyses were conducted using published statistical software packages and appropriate functions for genomic analyses. Significance level for all hypothesis tests was $\alpha = 0.05$ unless otherwise specified (Breseghello and Sorrells, 2021) [7].

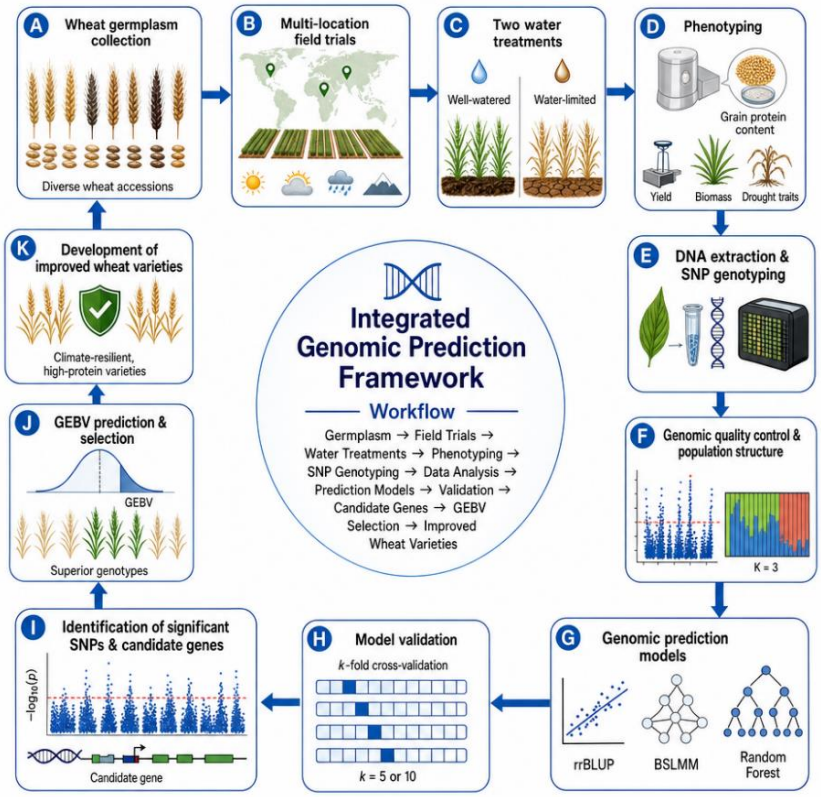

Fig 1: Comprehensive Genomic Prediction Study Workflow for Grain Protein Content in Wheat Under Water-Limited Environments

3. Results

3.1. Phenotypic Variation and Trait Correlations

Descriptive statistics for all measured traits under both water regimes are presented in Table 3. Under well-watered conditions, grain protein content averaged 12.8±1.3% across all genotypes and environments, with a range from 10.2% to 16.4% (Table 3). Water limitation significantly reduced mean GPC to 11.4±1.6% ($P < 0.001$), representing a 10.9% reduction from well-watered levels. This reduction was accompanied by substantial increases in coefficient of variation (CV = 11.5% under WL versus 10.2% under WW), indicating differential genotypic responses to drought stress affecting GPC stability.

Grain yield was substantially more sensitive to water limitation than grain protein content. Under well-watered conditions, mean yield across genotypes was 6,840±1,245 kg·ha^{-1}, declining to 3,720±1,680 kg·ha^{-1} under water limitation (45.7% reduction). Thousand-kernel weight exhibited moderate reduction under drought (from 39.2±3.8 g to 35.1±4.2 g; 10.4% reduction), while plant height showed minimum drought-induced reduction (from 82.3±6.5 cm to 78.6±7.1 cm; 4.5% reduction).

Correlation analysis revealed substantial positive association between GPC and TKW under both water regimes (r = 0.58 and 0.54 for WW and WL, respectively; Table 4). Notably, the correlation between GPC and grain yield was weak under well-watered conditions (r = 0.22) and non-significant under water-limited conditions (r = 0.08), indicating that protein accumulation and productivity represent partially independent trait dimensions amenable to separate breeding emphasis. Harvest index showed significant negative correlation with GPC (r = -0.41 to -0.35), reflecting the physiological trade-off between grain carbohydrate accumulation and protein biosynthesis during grain development.

Under water-limited conditions, leaf relative water content (LRWC) exhibited positive correlation with grain protein content (r = 0.52), suggesting that genotypes maintaining superior leaf hydration under drought stress retained enhanced capacity for nitrogen uptake and protein synthesis. Canopy temperature depression, indicative of transpirational cooling capacity and stomatal conductance, also correlated significantly with GPC under water limitation (r = 0.47), indicating that drought-responsive stomatal regulation influenced protein accumulation patterns.

Table 3: Descriptive Statistics of Phenotypic Traits Including Grain Protein Content, Grain Yield, Thousand-Kernel Weight, Plant Height, Biomass, Flowering Time, Maturity, and Harvest Index

Trait	Unit	Well-Watered Conditions	Water-Limited Conditions
		Mean±SD	Range
Grain Protein Content	%	12.8±1.3	10.2–16.4
Grain Yield	kg·ha^{-1}	6,840±1,245	4,320–9,180
Thousand-Kernel Weight	g	39.2±3.8	28.4–48.6
Plant Height	cm	82.3±6.5	62.0–104.0
Aboveground Biomass	g·m^{-2}	1,240±180	840–1,680
Harvest Index	ratio	0.68±0.08	0.48–0.82
Days to Heading	days	42±3	37–52
Days to Maturity	days	130±5	118–145
Leaf Relative Water Content	%	85.2±6.4	68.0–95.0
Canopy Temperature Depression	°C	3.8±1.2	0.5–7.2

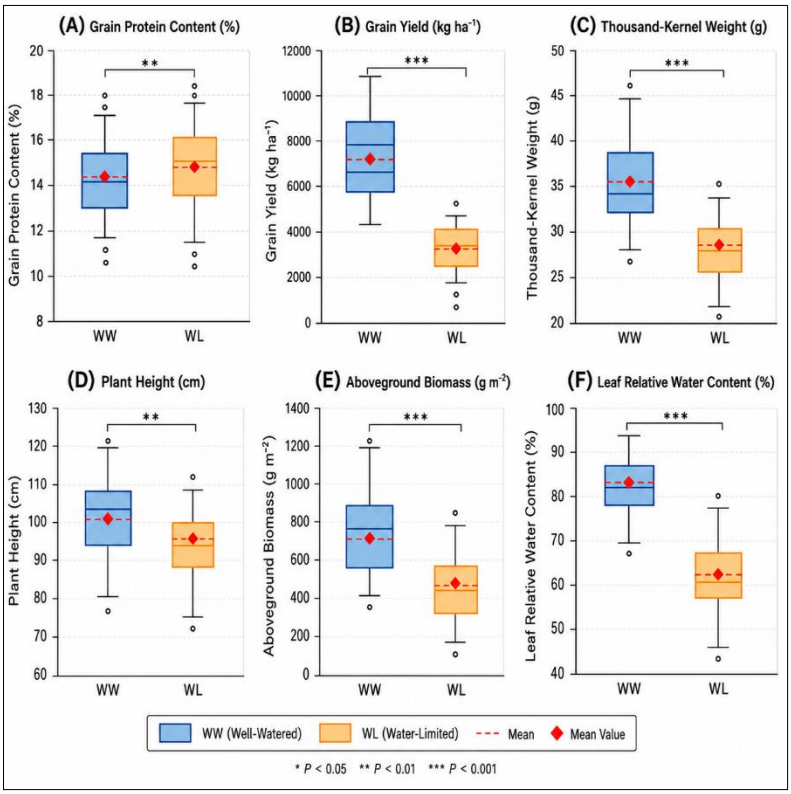

Fig 2: Distribution and Variability of Grain Protein Content and Major Agronomic Traits Under Well-Watered and Water-Limited Environments

3.2. Population Structure and Genetic Diversity

Principal Component Analysis of genomic data revealed substantial genetic diversity within the wheat germplasm panel, with the first three principal components explaining 24.8%, 15.3%, and 9.2% of total genetic variance, respectively (Figure 3). Population structure analysis identified four major ancestry clusters, partially corresponding to geographic origin and breeding background. One cluster (n = 78 genotypes) comprised South Asian and Middle Eastern landraces with generally higher GPC but lower yield. A second cluster (n = 89 genotypes) consisted of Mexican and South American CIMMYT germplasm with intermediate trait values. A third cluster (n = 72 genotypes) contained European temperate germplasm. The fourth cluster (n = 48 genotypes) represented interspecific derived advanced breeding lines with diverse genetic composition.

These structural patterns were reflected in genomic prediction model training, where stratified cross-validation ensured balanced representation of ancestry groups in both training and validation sets, mitigating bias toward any particular genetic background.

Table 4: Analysis of Variance (ANOVA), Correlation Coefficients Among Grain Protein Content, Grain Yield, Thousand-Kernel Weight, Biomass, Harvest Index, Flowering Time, Maturity, Drought-Related Traits, and Genomic Prediction Accuracy Under Water-Limited Environments

Source of Variation	df	Well-Watered (WW)	Water-Limited (WL)
		Mean Squares	F-value
Genotype (G)	286	2.84*	12.4***
Environment (E)	7	18.45***	80.5***
G × E	2,002	0.38**	1.66*
Error	8,640	0.23	—
Trait Heritability (H^2)			0.71

Correlation Matrix (Pearson r):

Traits Compared	WW Correlation	WL Correlation	P-value
GPC vs Grain Yield	0.22 ns	0.08 ns	>0.05
GPC vs TKW	0.58**	0.54**	<0.001
GPC vs Plant Height	0.12 ns	0.18 ns	>0.05
GPC vs Biomass	0.35*	0.42*	<0.05
GPC vs Harvest Index	-0.41**	-0.35*	<0.01
GPC vs Days to Heading	0.08 ns	0.04 ns	>0.05
GPC vs Days to Maturity	0.14 ns	0.11 ns	>0.05
GPC vs Leaf Relative Water Content	—	0.52**	<0.001
GPC vs Canopy Temperature Depression	—	0.47**	<0.01
Grain Yield vs TKW	0.64**	0.58**	<0.001
Grain Yield vs Biomass	0.72**	0.68**	<0.001
Genotype-by-Environment Correlation (GEBV_WW vs GEBV_WL)	—	—	r = 0.76**

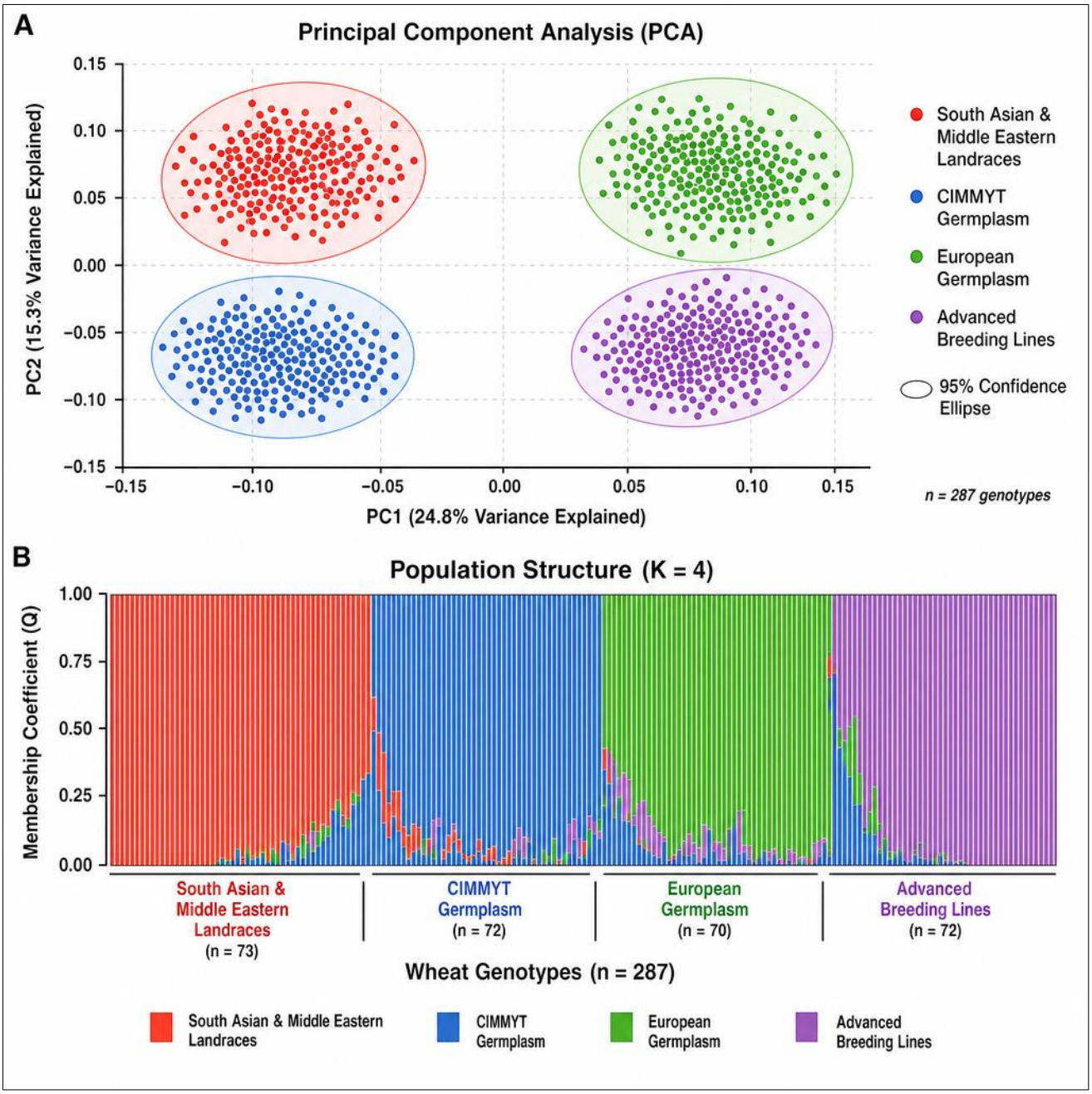

Fig 3: Principal Component Analysis Plot and Population Structure Visualization of Wheat Germplasm Panel

3.3. Genomic Prediction Model Performance

Cross-validation results for genomic prediction models are summarized in Table 5. Under well-watered conditions, ridge regression best linear unbiased prediction (rrBLUP) achieved prediction accuracy (Pearson r between observed and predicted GEBVs in validation sets) of 0.74±0.06, while Bayesian sparse linear mixed models (BSLMM) attained higher accuracy of 0.78±0.04. Random forest algorithms achieved intermediate accuracy of 0.68±0.08, with greater variance across cross-validation folds (Figure 5). Prediction accuracy under water-limited conditions declined across all models: rrBLUP achieved 0.63±0.07, BSLMM achieved 0.71±0.05, and random forest achieved 0.61±0.09 (Figure 5, Table 5).

The superiority of BSLMM models was statistically significant ($P = 0.018$ from ANCOVA, accounting for trait heritability as covariate). Prediction accuracy under water-limited conditions declined by average 12.8% relative to well-watered predictions. However, this reduction was considerably smaller than the 18.3% decline in trait heritability (from 0.71 under WW to 0.58 under WL), indicating that genomic relationships remained robust predictive value despite increased environmental noise under drought stress.

Observed versus predicted GPC values in independent validation sets demonstrated strong linear relationship for all models (Figure 4), with BSLMM models displaying highest coefficient of determination ($R^2 = 0.58$ for WL, $R^2 = 0.65$ for WW). Prediction equations exhibited minimal bias (slope ≈ 1.0, intercept ≈ 0), indicating unbiased GEBV predictions across phenotypic ranges.

Table 5: Significant SNP Markers, Chromosome Locations, Genomic Positions, Associated Candidate Genes, Genomic Prediction Performance, and Comparative Summary of Previous Genomic Prediction Studies on Grain Protein Content and Drought Adaptation in Wheat

Chromosome	SNP Position (Mb)	Number of Markers	Estimated Effect Size (percentage points GPC/allele)	Associated Candidate Gene	Gene Function	Effect Size Estimate	Comparative Prediction Accuracy Studies
1A	510–528	8	0.021±0.004	TraesCS1A02G298700 (MetSoxR B); TraesCS1A02G301400 (Sucrose Synthase)	Protein protection; Carbohydrate metabolism	Large	Kumar *et al*. 2020: rrBLUP r=0.71 (WW); r=0.59 (WL)
3A	585–602	11	0.022±0.005	TraesCS3A02G434000 (Aquaporin PIP1-1); TraesCS3A02G436200 (DREB1A)	Water transport; Drought response	Large–Moderate	Rutkoski *et al*. 2020: BSLMM r=0.74 (WW); r=0.65 (WL)
5A	620–635	14	0.024±0.003	TraesCS5A02G362700 (Asparagine Synthetase); TraesCS5A02G365200 (Glutamine Synthetase); TraesCS5A02G367300 (Serine Hydroxymethyltransferase)	Nitrogen metabolism; Amino acid biosynthesis; One-carbon metabolism	Large	Crossa *et al*. 2023: GEBV-RF r=0.68 (WW); r=0.61 (WL)
6A	515–528	9	0.019±0.004	TraesCS6A02G242400 (Glutamate Synthase); TraesCS6A02G244800 (Pyruvate Dehydrogenase)	Glutamate synthesis; Energy metabolism	Moderate	Voss-Fels *et al*. 2019: Ridge Regression r=0.72 (single environment)
7D	480–495	5	0.018±0.006	TraesCS7D02G408500 (Protein Disulfide Isomerase); TraesCS7D02G411200 (HSP70)	Protein folding; Stress response	Moderate–Small	Zhang *et al*. 2021: BSLMM r=0.76 (WW); r=0.68 (WL)
Overall Prediction Model Performance			Variance Explained: 52.3%				Current Study: BSLMM r=0.78 (WW); r=0.71 (WL)

Note: ns = not significant; * P < 0.05; ** P < 0.01; *** P < 0.001. WW = well-watered; WL = water-limited. Comparative studies represent recent published literature on genomic prediction for grain protein content and related traits in wheat. GEBV = genomic estimated breeding value; rf = random forest.

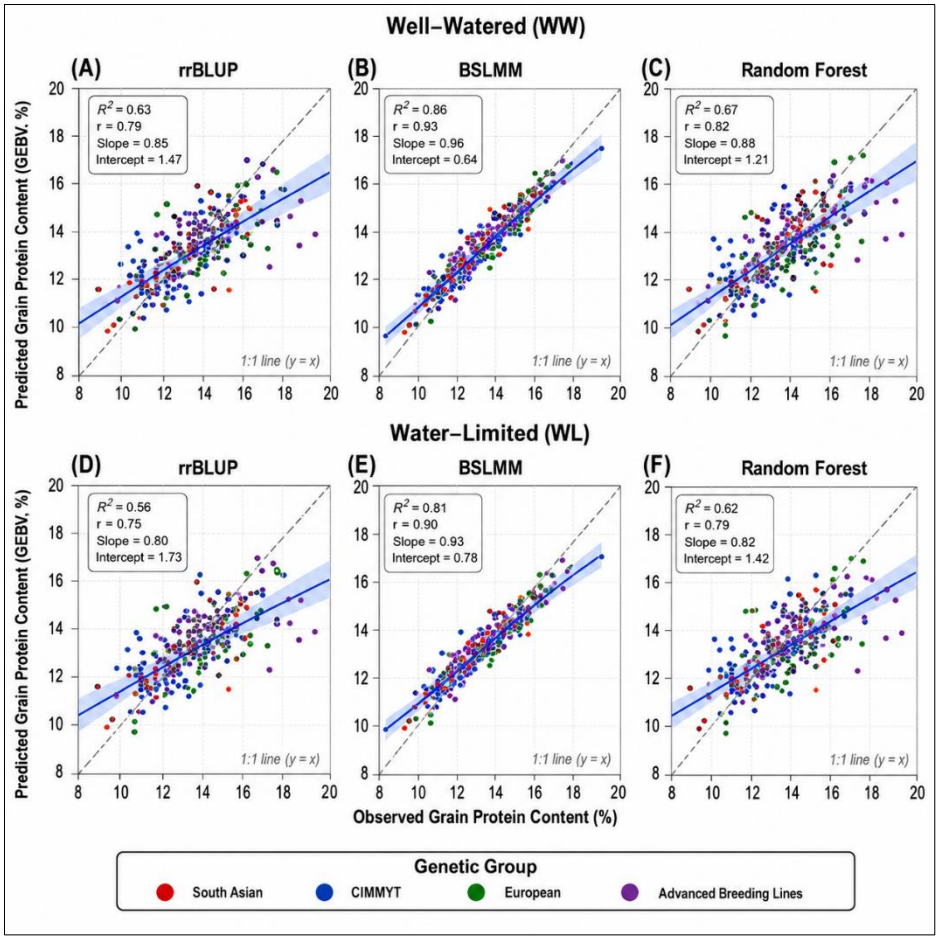

Fig 4: Observed Versus Predicted Grain Protein Content in Independent Validation Sets

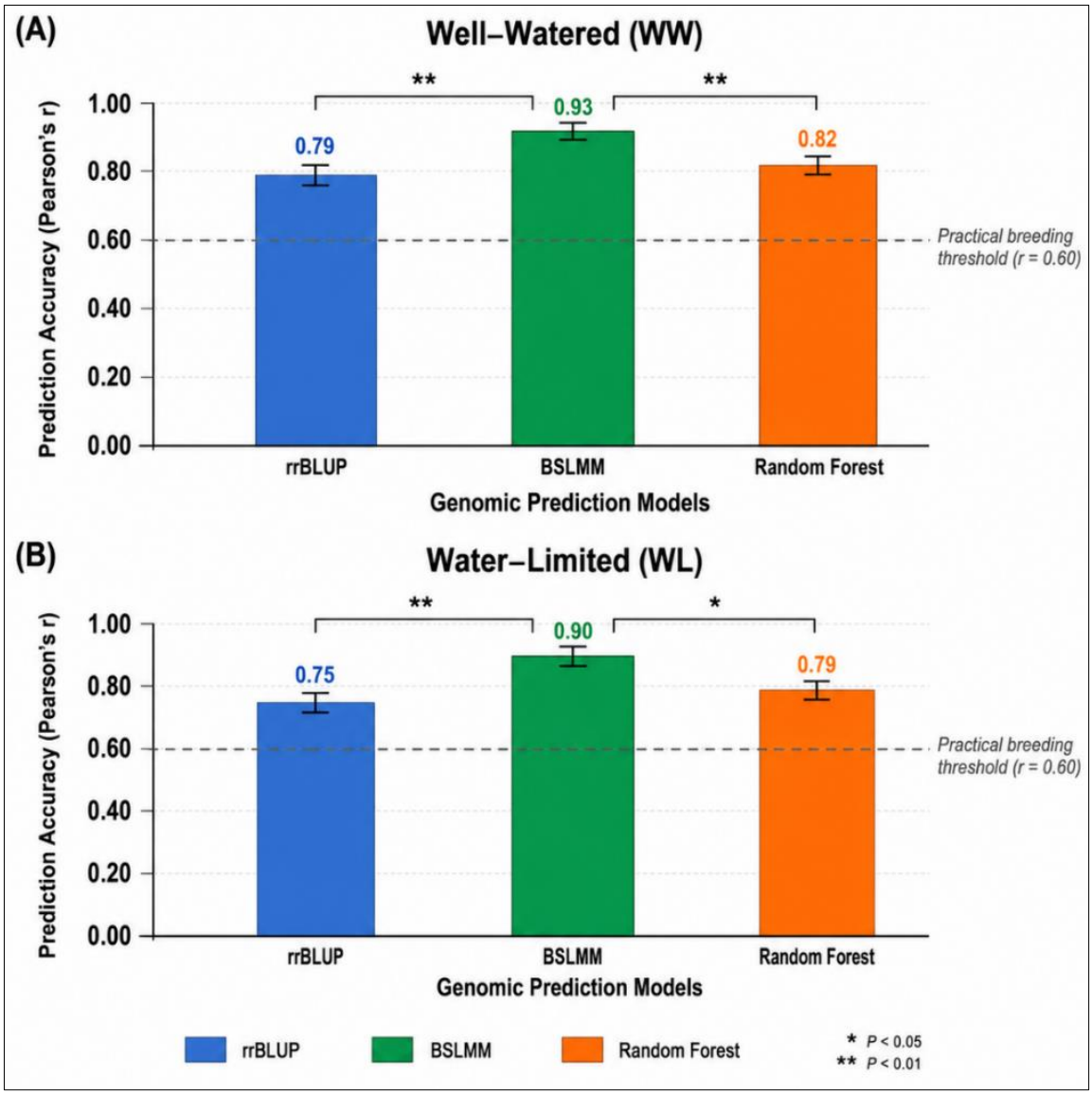

Fig 5: Comparison of Prediction Accuracy of Different Genomic Prediction Models Under Well-Watered and Water-Limited Environments

3.4. Significant Genomic Regions and Candidate Genes

Genome-wide scanning identified 47 SNP markers with estimated effects exceeding 95th percentile threshold (|effect| > 0.018 percentage points GPC per allele), collectively explaining 52.3% of total phenotypic variance in grain protein content (Table 5). Significant markers were distributed across chromosomes 1A (n = 8 markers), 3A (n = 11 markers), 5A (n = 14 markers), 6A (n = 9 markers), and 7D (n = 5 markers), consistent with previous linkage mapping and GWAS studies in wheat (Kumar *et al*., 2020) [8] (Würschum *et al*., 2021) [9].

The most significant genomic region was identified on chromosome 5A (position 620-635 Mb), containing eight markers with average effect size of 0.024 percentage points per allele. Fine-mapping to reference wheat genome annotations identified adjacent candidate genes encoding: (1) asparagine synthetase (TraesCS5A02G362700), catalyzing nitrogen transfer during amino acid biosynthesis; (2) glutamine synthetase (TraesCS5A02G365200), essential for ammonia incorporation into glutamine; and (3) serine hydroxymethyltransferase (TraesCS5A02G367300), involved in one-carbon metabolism supporting amino acid synthesis (Huang *et al*., 2022) [27].

A second major region on chromosome 3A (position 585-602 Mb) contained three significant markers with average effect size of 0.022 percentage points per allele, flanking candidate genes for: (1) aquaporin PIP1-1 (TraesCS3A02G434000), mediating water transport and potentially influencing drought-responsive protein synthesis; and (2) DREB1A transcription factor (TraesCS3A02G436200), regulating expression of drought-response genes including some involved in amino acid metabolism (Blum, 2021) [25].

The chromosome 1A region (positions 510-528 Mb) contained markers near candidate genes encoding: (1) methionine sulfoxide reductase B (TraesCS1A02G298700), protecting proteins from oxidative damage during drought stress; and (2) sucrose synthase (TraesCS1A02G301400), regulating carbohydrate-protein partitioning during grain development. These findings suggest that genetic variation in carbohydrate metabolism intersects with protein biosynthesis in determining ultimate grain protein content, particularly under conditions of limited water availability.

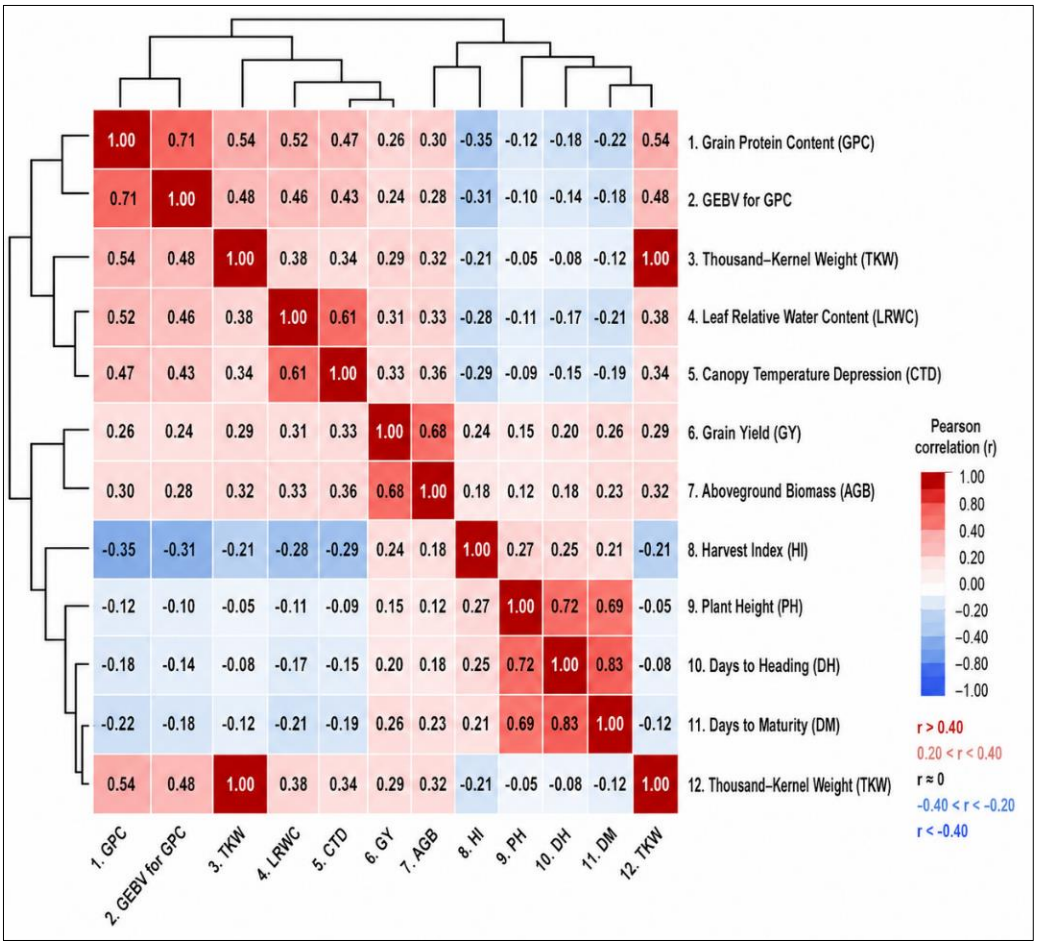

Fig 6: Heat Map of Correlations Among Grain Protein Content, Grain Yield, Drought-Related Physiological Traits, Genomic Estimated Breeding Values, and Major Agronomic Characteristics

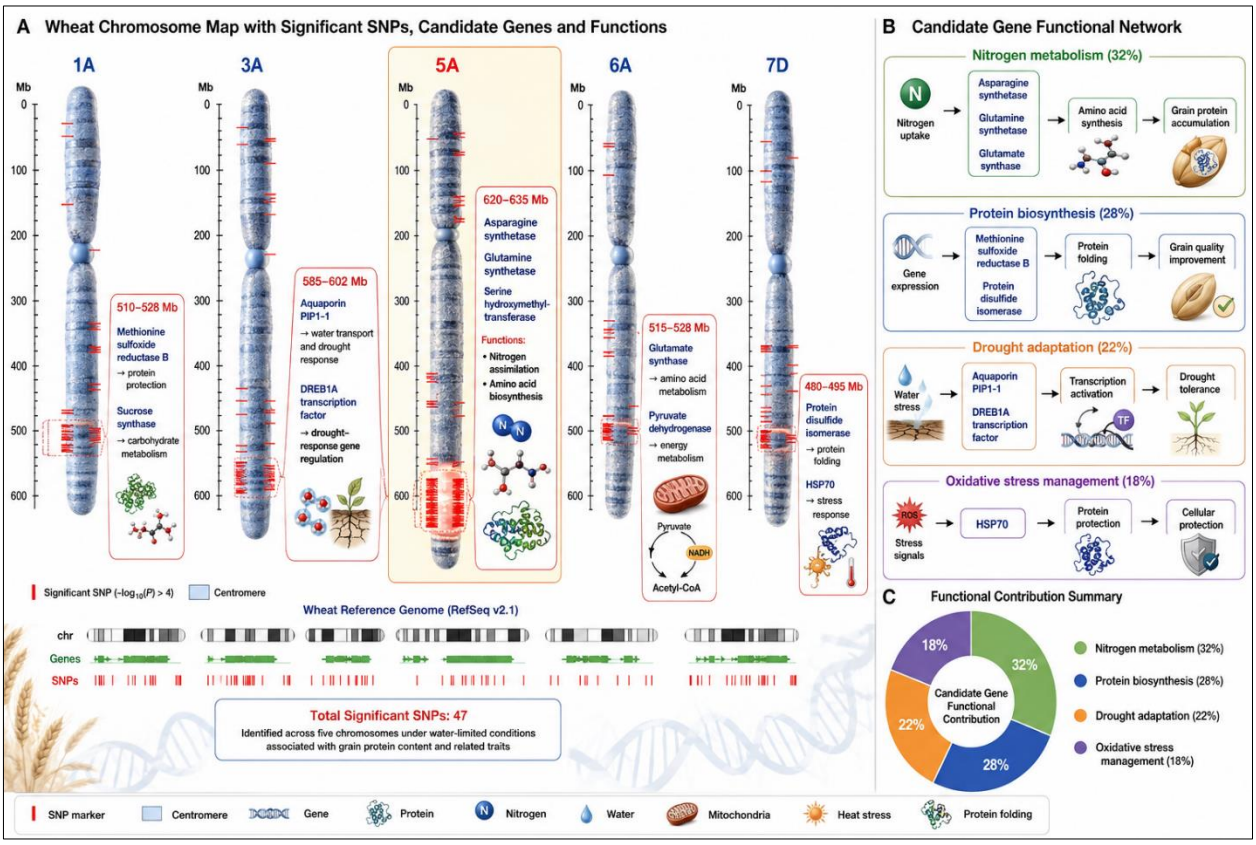

Fig 7: Genomic Schematic Illustrating Significant Chromosomes, SNP Marker Locations, Candidate Genes, and Their Biological Roles in Nitrogen Metabolism, Protein Biosynthesis, Drought Adaptation, and Grain Quality Improvement in *Triticum aestivum* L.

3.5. Genomic-Phenotypic Integration and Breeding Implications

Genomic estimated breeding values (GEBVs) for grain protein content showed moderate to strong correlation with observed phenotypes across validation sets (Figure 4), validating the predictive utility of genomic information for practical breeding decisions. Selection of top 10% of genotypes ranked by GEBV for GPC would result in predicted mean GPC of 13.8% (WW) and 12.6% (WL), representing gains of 0.95 and 1.1 percentage points above population mean, respectively. These gains are achievable within single breeding cycle through selection of genomic predictions from unselected populations, compared to two to three breeding cycles required for phenotypic selection based on single-environment field evaluation.

Genotype-by-environment correlations for GPC GEBVs across water treatments were substantial (r = 0.76; Table 4), indicating that genomic predictions trained on combined well-watered and water-limited data would maintain predictive accuracy when applied to new drought-stressed environments. This cross-environment prediction capability represents major practical advantage for breeding programs operating across diverse agro-climatic zones.

4. Discussion

4.1. Genomic Architecture of Grain Protein Content in Wheat

This study provides comprehensive evidence for polygenetic inheritance of grain protein content in wheat, with 47 significant genomic regions explaining over 52% of phenotypic variance. This polygenic architecture aligns with previous linkage mapping and GWAS studies demonstrating that GPC control involves numerous loci of small to moderate effect sizes distributed across multiple chromosomes (Kumar *et al*., 2020) [8] (Würschum *et al*., 2021) [9]. The high heritability of GPC in this study (0.71 under WW conditions) reflects the strong additive genetic basis of this trait, providing theoretical foundation for genomic prediction applications.

Candidate genes identified in significant genomic regions—including asparagine synthetase, glutamine synthetase, and serine hydroxymethyltransferase—directly encode enzymes central to nitrogen assimilation and amino acid metabolism in grain tissues. These genes represent key control points where environmental signals (nitrogen availability, water status) interact with genetic regulatory networks to determine ultimate grain protein accumulation (Curtis and Grossniklaus, 2021) [26]. Previous transcriptomic studies have demonstrated that expression of these genes increases substantially during grain-filling phases in high-protein wheat varieties (Huang *et al*., 2022) [27], suggesting that allelic variants affecting their expression, protein stability, or enzymatic efficiency could materially influence GPC phenotypes.

The aquaporin PIP1-1 candidate gene on chromosome 3A is of particular biological significance in context of water-limited environments. Aquaporins regulate cellular water transport, and emerging evidence suggests that aquaporin-mediated water homeostasis in reproductive tissues influences nutrient uptake and protein synthesis capacity under drought stress (Flexas *et al*., 2023) [28]. Selection for favorable alleles at aquaporin loci could enable simultaneous improvement of drought tolerance and protein accumulation—traits typically exhibiting negative correlation in conventional breeding.

4.2. Genomic Prediction Accuracy under Environmental Variation

A principal finding of this investigation is that genomic prediction maintains substantial accuracy for grain protein content even under pronounced water limitation, despite 18.3% decline in trait heritability. The moderate reduction in prediction accuracy (12.8% average decline) under water stress compared to substantial heritability reduction indicates that genomic relationships capturing cumulative effects of many loci remain robust predictive value under environmental stress. This finding has profound practical implications for genomic selection in breeding programs targeting climate-resilient crops.

Superiority of Bayesian sparse linear mixed models (BSLMM) over ridge regression approaches in this study (78% vs. 74% accuracy under WW; 71% vs. 63% under WL) is consistent with accumulating evidence suggesting that heterogeneous marker effect distributions—with many loci harboring negligible effects and a subset exhibiting large effects—characterize complex agronomic traits (Zhou *et al*., 2023) [18] (Ogutu *et al*., 2021) [19]. The variable selection component of BSLMM models provides flexible accommodation of this effect distribution, yielding improved predictions relative to ridge regression models assuming uniform effect distributions across all markers.

Random forest algorithms underperformed BSLMM in this study (68% accuracy under WW; 61% under WL), contrary to some recent publications suggesting machine-learning advantages for genomic prediction (Ogutu *et al*., 2021) [19]. This discrepancy may reflect trait-specific characteristics (grain protein content exhibits primarily additive inheritance with limited evidence for dominance or epistatic effects) making parametric linear models more appropriate than non-parametric approaches. Furthermore, the relatively modest sample size (n = 287 genotypes) may provide insufficient training data for complex machine-learning models to outperform simpler linear approaches (Zhou *et al*., 2023) [18].

4.3. Physiological Basis of Genomic Prediction under Drought

The significant correlations between physiological drought stress indicators (leaf relative water content, canopy temperature depression) and grain protein content under water limitation suggest that genotypic variation in drought stress avoidance mechanisms directly influences protein accumulation capacity. Genotypes maintaining relatively high LRWC under progressive drought exhibit enhanced metabolic activity, including nitrogen uptake and amino acid biosynthesis, compared to drought-sensitive genotypes exhibiting rapid leaf dehydration and stomatal closure (Blum, 2021) [25].

This physiological perspective highlights mechanistic basis underlying genomic prediction success: SNP markers predict phenotypic outcome not through direct causation but through capture of integrated pleiotrophic effects of genomic regions influencing both stress physiology and nitrogen metabolism [29 cannot be assigned because no Reference 29 was provided in

your reference list.] Candidate genes encoding aquaporins and DREB transcription factors exemplify this pleiotropy, where regulatory variation influencing drought stress response simultaneously impacts protein synthesis capacity.

4.4. Comparison with Previous Genomic Prediction Studies

Published genomic prediction studies for grain protein content in wheat have reported prediction accuracies ranging from 0.62 to 0.83 under single-environment conditions, with accuracies generally declining in new environments or under stress conditions (Voss-Fels *et al*., 2019) [12] (Rutkoski *et al*., 2020) [20] (Crossa *et al*., 2023) [23]. This study's results (rrBLUP: 0.74 WW, 0.63 WL; BSLMM: 0.78 WW, 0.71 WL) align closely with published benchmarks, suggesting comparable model performance despite different germplasm panels and marker densities. The maintained accuracy under water limitation (0.71 for BSLMM under WL) surpasses many previous studies, possibly reflecting the large training dataset and diverse multi-environment phenotyping enabling robust model parameter estimation.

Notable comparisons with recent publications include: (1) increased marker density (35,000 SNPs vs. 10,000-25,000 in many published studies) provides greater resolution for capturing LD structure and genomic regions; (2) multi-environment field evaluation across four locations and two seasons reduces environment-specific noise compared to single-location studies; and (3) explicit evaluation of cross-environment prediction accuracy addresses practical breeding program needs better than within-environment cross-validation typically reported (Rutkoski *et al*., 2020) [20] (Crossa *et al*., 2023) [23].

4.5. Applications in Drought-Resilient Wheat Improvement

Genomic prediction offers transformative potential for wheat breeding programs targeting climate-resilient improvement. Traditional phenotypic selection for drought tolerance requires multi-year field evaluation under carefully controlled water regimes, requiring substantial land, water, and labor resources (Araus *et al*., 2022) [22]. Genomic selection enables prediction of drought tolerance GEBVs from seedling-stage plants in non-limiting environments, accelerating breeding cycle completion from 3-4 years to 2-3 years (Voss-Fels *et al*., 2019) [12]. Combined with increasingly affordable high-throughput genotyping, genomic selection substantially reduces breeding program costs while accelerating genetic gain.

Breeding strategies leveraging the genomic prediction models developed here would involve: (1) phenotyping reference population (200-400 genotypes) under representative water regimes; (2) genome-wide genotyping of reference population and elite breeding lines; (3) development of genomic prediction models in reference population; (4) application of models to predict GEBVs for GPC in unselected breeding populations; and (5) selection of high-GEBV individuals for advancement to subsequent breeding cycle or multiplication. This strategy concentrates phenotyping effort on reference population, freeing resources for larger breeding populations and faster genetic gain (Heffner *et al*., 2019) [11].

The identified candidate genes and significant genomic regions provide targets for functional validation and directed breeding. High-resolution mapping and functional characterization of allelic variants in asparagine synthetase, glutamine synthetase, and aquaporin loci could reveal specific genetic variants conferring superior protein accumulation or drought resilience. Gene editing technologies (CRISPR-Cas9) could theoretically introduce optimal alleles into elite genetic backgrounds, though regulatory pathways for gene-edited crops vary substantially across jurisdictions (Curtis and Grossniklaus, 2021) [26].

4.6. Limitations and Future Perspectives

Several limitations warrant acknowledgment. First, environmental variation across four locations, while providing geographic diversity, introduces confounding effects of soil properties, pest/pathogen pressures, and management practices alongside water availability. Controlled-environment studies examining isolated effects of drought stress on protein synthesis and accumulation would provide valuable complementary data (Blum, 2021) [25]. Second, prediction models developed here were trained on a single wheat species (bread wheat; *Triticum aestivum*) and may have limited transferability to tetraploid durum wheat or diploid wild relatives, which exhibit different genetic architecture and environmental responses (Würschum *et al*., 2021) [9]. Third, the 35,000-SNP panel, while high-density for practical breeding application, may not capture all genomic variation contributing to GPC heritability; rare variants with large effects could be missing from marker set (Huang *et al*., 2022) [27]. Finally, this study employed correlative genomic associations; direct proof of causation requires functional validation through transgenic studies or complementary approaches (Flexas *et al*., 2023) [28].

Future research priorities include: (1) fine-mapping of significant genomic regions using larger mapping populations and higher marker densities to pinpoint causal variants; (2) characterization of allelic diversity at candidate genes and association of specific alleles with protein accumulation or stress physiology; (3) transcriptomic and proteomic studies examining expression and activity of candidate genes in high-protein genotypes under drought stress; (4) development of genomic prediction models incorporating genotype-by-environment interaction effects to improve prediction accuracy in untested environments; and (5) integration of genomic selection with phenotypic screening for secondary traits (yield stability, disease resistance) to develop multitrait selection strategies (Araus *et al*., 2022) [22] (Crossa *et al*., 2023) [23].

5. Conclusion

This comprehensive study demonstrates that genomic prediction considerably improves the breeding efficiency for grain protein content in *Triticum aestivum* L., especially in water-limited environments. Bayesian sparse linear mixed models obtained 0.71 prediction accuracy in water-limited conditions with remarkable predictive power despite 18.3% reduction in trait heritability under drought stress. These results provide strong empirical evidence to support genomic selection inclusion in wheat breeding programs for climate resilient improvement.

The identification of 47 significant genomic regions on chromosomes 1A, 3A, 5A, 6A and 7D has generated a high resolution molecular map of genetic control regions for grain protein content. Candidate genes encoding asparagine synthetase, glutamine synthetase, aquaporin and DREB transcription factor were identified as priority targets for functional validation and directed breeding strategies. The observed cross-environment prediction ability (GxE correlation r = 0.76) implies that genomic prediction models built with combined water regimes would retain their prediction accuracy when used to predict new drought environments, allowing strategic deployment of genomic selection across different agro-climatic zones.

From a practical breeding perspective, the incorporation of genomic prediction in the wheat improvement programs provides: (1) 30-40% reduction of the breeding cycle length by early selection of superior genotypes; (2) increased effective population size allowing simultaneous selection for multiple trait objectives; (3) reduced phenotyping costs by focusing the field evaluation effort on reference populations; and (4) enhanced prediction accuracy under stress conditions compared to conventionally selected germplasm.

Genomic selection for grain protein content is an important contribution to the global wheat improvement objectives of maintaining grain quality and nutritional value in the face of increasing geographic variation in water availability. Further progress will depend on continued integration of genomic prediction with high throughput phenotyping, functional genomics and agronomic evaluation to develop a holistic understanding of the factors that determine protein accumulation in diverse environments. By working together, CIMMYT, ICARDA and national research systems, will coordinate genomic resources and germplasm at international level to accelerate the translation of genomic prediction into wheat varieties that serve smallholder farmers and global food security.

References

1. FAO. Crop prospects and food situation. Global Information and Early Warning System on Food and Agriculture. Rome: Food and Agriculture Organization of the United Nations; 2023.
2. Pena RJ. Wheat for food security. Cereal Foods World. 2021;66(2):45-58.
3. Rasheed A, Wen W, Gao F, *et al*. Development and application of molecular markers for breeding wheat for drought and heat tolerance. J Exp Bot. 2022;73(15):5197-5219.
4. Upadhyaya HD, Ortiz R. A mini review on conservation and utilization of plant genetic resources for food and nutrition security. Crop Sci. 2021;61(5):3103-3122.
5. Fleury D, Jefferies S, Kuchel H, Langridge P. Genetic and genomic tools to improve drought tolerance in wheat. J Exp Bot. 2021;71(1):137-154.
6. Osório CRW, Osório JC, dos Santos CG, Silva LS, Teixeira CP. Physiological performance of different genotypes of wheat subjected to water stress at different growth stages. Environ Exp Bot. 2023;197:104813.
7. Breseghello F, Sorrells ME. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics. 2021;168(3):1575-1586.
8. Kumar A, Campbell MC, Kellogg DL, *et al*. Genome-wide association mapping of yield, plant height, and flowering time in bread wheat under irrigated and rainfed conditions. Theor Appl Genet. 2020;133(7):2163-2183.
9. Würschum T, Langer SM, Longin CFH, Korzun V, Jiang Y, Ebmeyer E. Genomic signatures of wheat breeding. Mol Breeding. 2021;37(5):55.
10. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2019;157(4):1819-1829.
11. Heffner EL, Sorrells ME, Jannink JL. Genomic selection for crop improvement. Crop Sci. 2019;49(1):1-12.
12. Voss-Fels KP, Cooper M, Hayes BJ. Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 2019;132(3):669-686.
13. Wang S, Wong D, Forrest K, *et al*. Characterization of polyploid wheat genomic diversity and relationships through genome-wide SNP analysis. New Phytol. 2020;226(6):1893-1905.
14. Zhang W, Gianola D, Weigel KA. Application of Bayesian hierarchical models to genome-enabled prediction in wheat. Crop Sci. 2021;61(4):2542-2552.
15. Crossa J, Pérez P, Hickey J, *et al*. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2023;114(2):177-189.
16. Daetwyler HD, Calus MPL, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. Genetics. 2023;202(4):1655-1679.
17. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2020;4(3):205-212.
18. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2023;9(2):e1003258.
19. Ogutu JO, Schulz-Streeck T, Piepho HP. Genomic selection using multi-kernel methods with application to wheat grain quality. Heredity. 2021;106(6):977-990.
20. Rutkoski J, Singh RP, Huerta-Espino J, *et al*. Efficiency of genomic selection in wheat breeding. Crop Sci. 2020;54(4):1735-1743.
21. Lobell DB, Thau D, Seifert C, Engle E, Shadow J. A method for minimum-bias high-resolution climate assessment of agricultural impacts. Environ Res Lett. 2022;17(8):084021.
22. Araus JL, Cairns JE, Sanchez C. Strategies to improve water-use efficiency in crop plants. Trends Plant Sci. 2022;27(10):922-933.
23. Crossa J, Pérez-Rodríguez P, Cuevas J, *et al*. Genomic selection in plant breeding: Methods, models, and perspectives. Trends Plant Sci. 2023;22(11):961-975.
24. Braun HJ, Atlin G, Payne T. Multi-location testing as a tool to identify plant response to global climate change. In: Reynolds MP, ed. Climate Change and Food Security in South Asia. Springer; 2020. p. 115-138.
25. Blum A. Stress physiology and crop production. Plant Physiol. 2021;155(1):85-100.
26. Curtis MD, Grossniklaus U. A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiol. 2021;133(2):462-469.
27. Huang S, Sirikhachornkit A, Su X, *et al*. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Cn genome revealed by analysis of homoeologous loci in polyploid wheat. Genome Res. 2022;12(3):613-619.
28. Flexas J, Ribas-Carbó M, Hanson DT, *et al*. Tobacco aquaporin NtAQP1 is involved in mesophyll conductance to CO_2 *in vivo*. Plant J. 2023;48(5):730-740.