scientific reports

Check for updates

OPEN

Genomic prediction for agronomic traits in a diverse Flax (*Linum usitatissimum* L.) germplasm collection

Ahasanul Hoque[1,2], James V. Anderson[3] & Mukhlesur Rahman[1]✉

Breeding programs require exhaustive phenotyping of germplasms, which is time-demanding and expensive. Genomic prediction helps breeders harness the diversity of any collection to bypass phenotyping. Here, we examined the genomic prediction's potential for seed yield and nine agronomic traits using 26,171 single nucleotide polymorphism (SNP) markers in a set of 337 flax (*Linum usitatissimum* L.) germplasm, phenotyped in five environments. We evaluated 14 prediction models and several factors affecting predictive ability based on cross-validation schemes. Models yielded significant variation among predictive ability values across traits for the whole marker set. The ridge regression (RR) model covering additive gene action yielded better predictive ability for most of the traits, whereas it was higher for low heritable traits by models capturing epistatic gene action. Marker subsets based on linkage disequilibrium decay distance gave significantly higher predictive abilities to the whole marker set, but for randomly selected markers, it reached a plateau above 3000 markers. Markers having significant association with traits improved predictive abilities compared to the whole marker set when marker selection was made on the whole population instead of the training set indicating a clear overfitting. The correction for population structure did not increase predictive abilities compared to the whole collection. However, stratified sampling by picking representative genotypes from each cluster improved predictive abilities. The indirect predictive ability for a trait was proportionate to its correlation with other traits. These results will help breeders to select the best models, optimum marker set, and suitable genotype set to perform an indirect selection for quantitative traits in this diverse flax germplasm collection.

Flax (*Linum usitatissimum* L.), a natural source of oil and fiber, has been grown throughout the world since prehistoric times and has considerable economic importance[1]. Flaxseed plays an important role in human nutrition[2,3] by providing oil rich in omega-3 fatty acid, dietary fibers, and anti-carcinogenic lignans. Flaxseed oil is used for different industrial purposes such as ink making, varnishing, painting, and road carpeting due to its specific drying properties[4]. Additionally, flaxseed meal has value as poultry and animal feeds[5,6] while flax fiber is used in making linen cloth and different bio-industrial products[7].

Among the flax-growing countries, Kazakhstan produces the most oilseed flax followed by the Russian Federation, Canada, China, and the USA, whereas France alone produces three-fourths of fiber flax in the world followed by Belgium, Belarus, Russian Federation, and China[8]. In the United States, North Dakota (ND) has the greatest % of cultivated flax acres, which covers about 71% (215 million acres) and 80% (3.75 billion bushels) of U.S. flax acreage and production, respectively, and annually contributes about $46 million U.S. dollars to the national economy (Data averaged from 2017 to 2021)[9]. Like many other cultivated crops, flax production in ND is being challenged by different biotic and abiotic stresses[10,11]. To combat these challenges and meet farmers' desires for high-yielding varieties with increased oil and protein content, North Dakota State University (NDSU) runs a moderate-size flax breeding program. The program utilizes classical breeding methods, especially modified bulk methods to develop varieties, which is expensive, laborious, and time-consuming. To speed up the breeding process, it is prime time to adopt cutting-edge breeding tools such as marker-assisted selection (MAS) and genomic selection (GS) in the program.

[1]Department of Plant Sciences, North Dakota State University, Fargo, ND, USA. [2]Department of Genetics and Plant Breeding, Bangladesh Agricultural University, Mymensingh 2202, Bangladesh. [3]USDA-ARS, Edward T. Schafer Agricultural Research Center, Fargo, ND, USA. ✉email: md.m.rahman@ndsu.edu

Initially, breeders utilized marker-trait associations revealed by linkage mapping and genome-wide association mapping for MAS in breeding programs to enhance efficiency and genetic gain[12]. To date, MAS has been successfully used to improve monogenic or oligogenic traits in many major crops such as rice[13,14], wheat[15–17], maize[18–20], etc. However, the improvement of quantitative traits controlled by multiple QTLs with minor effects is challenging. A multi-marker MAS system can be used to improve quantitative traits, but it is very difficult to identify and account for all the allele effects[21,22]. Breeders can overcome the limitations of MAS by using a genome-wide selection approach. Genomic selection (GS), also known as genomic prediction, by considering all marker effects regardless of significance holds a promise to accelerate the rate of genetic gain in the case of quantitative traits[23,24]. The GS was first successfully used in animal breeding, where it was applied to dairy cattle[25]. In recent days, the low genotypic cost compared to phenotypic cost has made the GS an attractive decision tool to select and evaluate accessions in diverse germplasm collections.

To date, many statistical models for GS have been developed. Initially, the RRBLUP model was widely used[23,26]. Later, plethora of linear or parametric models[27–30] and non-linear or non-parametric models[31–33] have evolved. Linear models capture only additive gene action effects, whereas non-linear models capture additive and non-additive (dominance, epistasis, and pleiotropy) interactions. Different models vary in their underlying assumptions and algorithms and many authors confirmed that no single model worked best across traits, rather a particular model outperformed for a particular trait[34–37]. That is why it is always recommended to test multiple models across traits to achieve maximum prediction accuracy. Along with models, various factors such as relatedness between the training set and validation set[38,39], correlation among studied traits[40,41], trait heritability, marker density, QTL size with effects[34,42–46] and genetic diversity of studied collection[36,47,48] also affect the prediction accuracy.

Plant breeders have successfully utilized GS to accelerate the varietal development process in rice[49–52], wheat[53–57], maize[58–61] and other crops. Despite the wide application of GS in many crops, its utilization in flax breeding has not flourished yet. Because very few reports are available regarding the utilization of GS in biparental[62,63] and diverse[64] flax populations, there is great potential for evaluating GS in diverse flax collections. This study aims to (1) investigate the feasibility of implementing genomic prediction for various agronomic traits, (2) identify the most effective prediction models and optimal marker numbers to maximize predictive abilities across traits, and (3) assess how marker-trait associations, population structure, and trait correlations influence predictive abilities for diverse traits.

Materials and methods

Plant materials

We collected 500 flax accessions and their wild relatives from the North Central Regional Plant Introduction Station (NCRPIS), Ames, Iowa, USA. All genotypes were grown in the field as single rows. We discarded the heterogeneous rows and kept the homogeneous lines for parental stock. Finally, we made a core collection of 337 flax germplasm accessions, which comprises homogeneous lines from NCRPIS, NDSU-released varieties and advanced breeding lines, and varieties developed by different institutes in the USA and Canada (Supplementary Table S1). The advanced breeding lines (F_7 generation) were obtained by crossing different parents in various combinations. The core collection is being maintained through selfing. NCRPIS and NDSU are public institutions that comply with all required regulations for utilizing seed materials for research and development purposes.

DNA extraction, sequencing, and SNP calling

Young leaves collected from 30-day old plants were used as the source of DNA. The collected leaf samples from each genotype were lyophilized and subsequently pulverized using stainless beads in a plate shaker. DNA was extracted from the ground leaf tissue using a Qiagen DNeasy Kit (Qiagen, CA, USA) according to the manufacturer's protocol. A NanoDrop 2000/2000c Spectrophotometer (Thermofisher Scientific) was used to measure the DNA concentrations. The GBS library was prepared using the ApekI enzyme[65] and sequencing of the library was accomplished using an Illumina HiSeq 2500 sequencer at the University of Texas Southern Medical Center, Dallas, Texas, USA. Identification of SNPs was based on a 120-base kmer length and minimum kmer count of ten using the TASSEL 5 GBSv2 pipeline[66] and the flax reference genome[67] (available at: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/224/295/GCA_000224295.2_ASM22429v2/). The reads were aligned to the reference genome using the Bowtie 2 (version 2.3.0)[68] alignment tool, which identified 243,040 SNPs that passed all required steps of the TASSEL 5 GBSv2 pipeline. Though flax is a strictly self-pollinating crop and inbred lines were used, there was a possibility of heterozygous SNPs due to artefactual collapse of homologous sites during alignment. We removed the heterozygous SNPs and filtered the row SNP set using VCFtools[69] following the criteria: minor allele frequency (MAF) ≥ 0.05, missing values (max-missing) $\leq 25\%$, depth (minDP) ≥ 3, min-alleles = 2 and max-alleles = 2. This filtering process yielded 26,171 bi-allelic high-quality SNP markers.

Phenotyping

We planted 337 genotypes following an augmented row-column design[70] with three standard checks (ND Hammond, Gold ND, and Omega), and the checks were diagonally placed to cover spatial heterogeneity (Supplementary Fig. S1). Each check was replicated 20 times per trial and experiments were conducted at Fargo, ND, USA (46.8772° N, 96.7898° W) for three consecutive years (2018, 2019, and 2020) and at Carrington, ND, USA (47.4497° N, 99.1262° W) in two consecutive years (2019, 2020). Hereafter, we referred to the location-year combinations as environments: E1 (Fargo, 2018), E2 (Fargo, 2019), E3 (Carrington, 2019), E4 (Fargo, 2020), and E5 (Carrington, 2020). In E1, we planted a single row (4 m long) per genotype, 7 gm of seeds per row due to a shortage of seeds. Later on, for all environments, for each genotype, we used 8 (4 m × 2 m) m^2 four-row plots and 30 gm of seeds per plot for planting. Standard fertilization and cultural practices were used throughout the

experiment. Data for nine agronomic traits was recorded in all environments and seed yield in four environments. For each genotype, data measured on different traits was based on the criteria and methods described by Nôžková et al. (2011)[71] with minor modifications. Days to flowering was determined as the number of days from planting to when approximately 50% of plants per plot start flowering. Plant height was measured as the length of the main stem at maturity from the hypocotyl ending point to the plant's top. The technical length was measured as the length of the main stem at maturity from the end of the hypocotyl to the point where branching starts. The branch number per plant was counted as the lateral branches of the main stem inflorescence. In this case, only the primary lateral branches were considered. The boll number per plant was the capsule number carried by the main stem inflorescence. Thousand seed weight was the weight of exactly 1000 seeds. Seed area, seed width, and seed length were calculated as an average of 1000 seeds. We used a MARVIN seed analyzer (GTA Sensorik GmbH) to measure seed-related traits. We harvested each plot separately and measured the grain weight in grams as yield per plot.

Phenotypic data analysis

In this study, a two-stage analysis of phenotypic data was performed. In stage I, the best linear unbiased estimates (BLUEs) and other statistics for all genotypes within each environment were determined using the following model:

$$y = X\tau + e_R + e_C \tag{1}$$

where X is the design matrix, τ is the fixed effect of genotype, e_R is the random effect of row and e_C is the random effect of the column.

In stage II, we fitted the BLUEs and weights from stage I analysis in Eq. (2) and estimated the best linear unbiased predictions (BLUPs) of genotypes across all environments.

$$y_{ij} = \mu + G_i + E_j + GE_{ij} + e_{ij} \tag{2}$$

where y_{ij} is the observed phenotypic value of the ith genotype in the jth environment, μ is the overall mean, G_i is the random effect of the ith genotype, E_j is the fixed effect of the jth environment, GE_{ij} is the G × E interaction term and e_{ij} is the residual error. The analysis was done using the R-shiny app MrBean (https://beanteam.shinyapps.io/MrBean/).

We calculated the heritability of each trait in each environment and combined all environments using the following formula proposed by Cullis et al. (2006)[72]:

$$H^2_{Cullis} = 1 - \left(\frac{PEV}{md * V_g}\right) \tag{3}$$

where the genotypic predicted error variance is PEV, V_g is the genotypic variance and md is mean values from the diagonal of the relationship matrix. The heritability calculation was done using the R package Sommer[73].

We calculated Pearson correlation among different traits within each environment and combined all environments using observed unadjusted phenotypic values. The correlation of phenotypic values of a trait observed in different environments was also calculated; for this purpose, we used the R package corrplot[74].

Structure and linkage disequilibrium (LD) analysis

An admixture model-based structure analysis of the whole germplasm set was conducted using STRUCTURE[75] software utilizing the whole SNP marker set (26,171). To strengthen the result, structure analysis was run at various combinations of burn-in lengths (5000–50,000) and Monte Carlo Markov Chain (MCMC) lengths (5000–100,000). Each combination was replicated 10 times per K (K1–K10). We used both the Delta K approach[76] and four alternative statistics[77] to identify the optimum number of clusters as the Delta K approach gives a variable number of clusters at different combinations of burn-in lengths and MCMC lengths[120,121]. StructureSelector[78] was used for this purpose. We assembled 10 replicates of the Q-matrix for the best-fitted cluster number using CLUMPP[79]. Principal component analysis (PCA) was run using a covariance standardized approach in TASSEL[80]. To show the genetic divergence among identified clusters, pairwise F_{st}[81] was calculated using Arlequin3.5[82] at 10,000 permutations. Gaussian finite mixture model-based clustering of the collection was fitted via the EM algorithm in R package mclust[83] using phenotypic data. We also visualized the PCA and phenotype-based clustering output using the ggplot2 R package[84].

Chromosome-wise LD (r^2 values) among SNPs was calculated using the 26,171 SNP markers in PopLDdecay software[85]. The LD decay rate was defined as a half-decay distance, at which observed r^2 between sites decays to less than half of the maximum r^2 value[86]. For this purpose, we wrote R scripts combining various R-packages (available on personal communication).

Genomic prediction models' comparison

In the case of genomic prediction, the linear model equation is unsolvable as the explanatory variables (marker number) exceed the observation number. Researchers solve this problem by using ridge regression or Bayesian computations or parametric method and penalized regression or semi-parametric method[87]. Generally, all methods are fitted to the basic skeleton (Eq. 4) with modifications:

$$y_{ij} = \mathbf{\mu} + Zu + \varepsilon \tag{4}$$

where y is the phenotypic value (BLUP), μ is the fixed intercept, Z is the marker matrix, μ is the marker effect vector and ε is a residual vector.

In this study, we assessed the predictive ability of different models for different traits using 26,711 SNP markers and BLUP values. We used 14 different parametric and semi-parametric models such as GBLUP[88–90], EGBLUP[91], RR[23,26,92], LASSO[93], EN[94], BRR[95], BA[23], BB[96], BC[97], BL[29], RKHS[98], RF[99], SVM[100,101] and MKRKHS[33]. Details of the models are available in previously published research articles[34,95]. For this purpose, we used the R package BWGS pipeline[34]. For each trait, we assessed the predictive ability as Pearson correlation between genomic estimated breeding values (GEBVs) and phenotypic values (BLUPs) of the validation set (VS). For this, a fivefold cross-validation approach was used, i.e., we randomly selected 80% of the collection as a training set (TS) and the remaining 20% as a validation set (VS). The process was repeated 100 times for each model and finally, the predictive ability was reported as average across 100 replicates. One-way ANOVA was done to explore whether the variation among predictive ability values by different models for each trait was significant or not. Then lettering was done by multiple comparison (Tukey) test to separate the averaged predictive ability values into groups. The model that gave the best predictive ability for a particular trait was declared as the best-fitted model for the corresponding trait. For subsequent analyses, we only used the best-fitted model identified in this stage for specific traits.

Marker subsets preparation

The predictive ability for each trait was assessed using different subsets of the whole marker set. The marker subsets were made based on linkage disequilibrium (LD) decay distance and random selection. We thinned the whole marker set using chromosome-wise LD decay distance, which yielded 5362 markers. For this purpose, we used chromosome-wise half-decay distance, at which observed r^2 between sites decays to less than half of the maximum r^2 value [86]. We also selected subsets of 20, 200, 1000, 3000, 7000, and 13,000 markers based on random sampling to minimize or avoid any biases. Using a five-fold cross-validation approach, the predictive ability for each subset was measured. The process was repeated 100 times and finally, predictive ability was reported as average across 100 replicates.

Marker subsets based on marker-trait association

Significant markers for each trait were grouped based on genome-wide association mapping (GWAS) results. This was done in five different ways. In the case of scenario-I, we conducted SNP-based GWAS for all traits within each environment (BLUEs) and combining all environments (BLUPs) using different single locus and multi-locus models such as general linear model (GLM)[102], mixed linear model (MLM)[103], compressed MLM (CMLM)[104], enriched compressed MLM (ECMLM)[105], settlement of MLM under progressively exclusive relationship (SUPER)[106], multiple loci MLM (MLMM)[107], fixed and random model circulating probability unification (FarmCPU)[108] and Bayesian information and linkage-disequilibrium iteratively nested keyway (BLINK)[109]. The R package GAPIT (version 3)[110] was used to run GWAS and the best-fitted model for a particular trait was determined based on the mean of squared difference (MSD) values and QQ plots. Using the best-fitted model output, we identified significant SNPs associated with a particular trait based on a p-value threshold. The p-value threshold was calculated by dividing the type-I error rate (α) by the effective number of independent tests ($Meff$) at $\alpha = 0.05$[111]. In this study, the p-value threshold was 0.000103. We grouped the significant markers for each trait considering each environment (BLUEs) and combined environment (BLUPs). In the case of scenario-II, GWAS was conducted using only an MLM model based on combined environment BLUP values. The 20 markers for each trait having the lowest p-value were grouped. Then the selected marker set from scenario-I & II were used to assess predictive ability 100 times using a five-fold cross-validation approach and predictive ability was reported as average across 100 replicates. In the case of scenario-III, we randomly divided the whole collection into TS (80% of the whole collection) and VS (20% of the whole collection) 15 times. Each time we did GWAS using TS only following the MLM[103] model and grouped the most significant 20 markers for each trait. Then, predictive ability was assessed and reported as the average across 15 replicates. In the case of scenario-IV, GWAS was conducted as one-way ANOVA using the R function *lm*, and every marker was tested one at a time using phenotype (BLUP values) considering the whole collection. The markers having a p-value less than 0.001, 0.01, and 0.05, respectively were grouped separately and then were used for predictive ability calculation 100 times with five-fold cross-validation. In the case of scenario-V, GWAS was done as one-way ANOVA using TS only to select the marker set having a p-value less than 0.001, 0.01, and 0.05. The selected marker set was then used to assess the predictive ability for each trait.

Predictive ability considering population structure

We investigated the confounding effects of population structure on predictive ability. To minimize the effect of population structure, we used genotypic subsets having less divergent genetic clusters and incorporated the Q-matrix from structure analysis output into the RR-BLUP model. Genotypic subsets were made by discarding the most divergent (clusters showing the highest pairwise F_{st} value to other clusters) genetic clusters P3 and P4 from the whole collection. Albeit a smaller sample size, we also assessed cluster-wise predictive ability. In all cases, we used five-fold cross-validation 100 times.

To explore the effect of population structure, we did stratified sampling to cover maximum genetic variance and calculated predictive ability following two different methods (M-I and M-II). The whole collection was arranged into small groups of 25, 50, 75, 100, 125,150, 175, 200, 250, and 300 genotypes by randomly selecting genotypes from each cluster proportional to the size of the cluster. In the case of method-I (M-I), 100 times five-fold cross-validation within each subset was done to report the predictive ability as the average across all runs. In the case of method -II (M-II), each subset was used as TS and the remaining genotypes as VS. For

this purpose, we made 20 replicates of each subset. The predictive ability was calculated as Pearson correlation between GEBVs and BLUP values of VS and was reported as an average across 20 times. There was overlapping of genotypes among replicates as genotypes were selected randomly from each cluster proportional to the size of the cluster each time.

Indirect predictive ability calculation

Indirect predictive ability for each trait, considering other traits separately, was calculated using the five-fold cross-validation schemes 100 times. For example, we estimated the GEBVs of the validation set for plant height. Then we calculated the indirect predictive ability for seed yield by plant height as Pearson correlation between GEBVs of the validation set based on plant height and BLUP values of that set considering seed yield. The same was repeated 100 times and predictive ability was reported as average across 100 replicates. The same procedure was followed for different trait combinations.

Results

Phenotypic variability

The genotype collection showed continuous variation for all traits under all environments (Supplementary Fig. S2). Under the five environmental conditions, days to flowering ranged from 36 to 67, plant height ranged from 20.0 to 82.8 cm, technical length ranged from 6.5 to 60.7 cm, branch number ranged from 2.7 to 12.3 and boll number ranged from 5 to 50. Among the traits, for the boll number, a maximum CV value of 36.2% was observed in E3. However, seed-related traits had relatively lower CV values compared to other traits. Among these, thousand seed weight had a maximum CV value of 16.0% in E1. Seed yield was evaluated under four environments, which ranged from 87 to 630 g per plot with a CV of 35.7 to 41.4%. Boll number and seed yield had the lowest heritability (< 0.40) compared to other traits under all environments. Days to flowering, branch number, and seed width indicated both low to high (0.18 to 0.93) environment-specific heritability, whereas it was high (> 0.60) for plant height, technical length, thousand seed weight, seed area, and seed length. All the traits exhibited high heritability (> 0.65) combining all environments except the branch number. The details of phenotypic variability are presented in Supplementary Table S2.

Phenotypic correlation

We investigated the phenotypic association among different traits and seed yield (Supplementary Fig. S3) in all environments. Days to flowering showed both positive and negative weak correlations with other traits across environments. Plant height, technical length, branch number, boll number, and seed yield had a positive significant correlation among them across environments. Among all combinations, the best positive significant correlation was found between plant height and technical length ($r > 0.77$), boll number and seed yield ($r = 0.69$ to 0.79) in all environments. Seed-attributing traits such as thousand seed weight, seed area, width, and length were significantly positively correlated with each other, but they exhibited mostly weak correlations with the remaining traits and seed yield across environments.

We also calculated the correlation among environments for each trait (Supplementary Fig. S4) and found comparatively low positive associations among environments for days to flowering and branch number, whereas it was good and positive for the remaining traits.

Population structure and linkage disequilibrium (LD) analysis

Structure analysis revealed 3 to 9 clusters based on the Delta K approach[76] and 4–5 clusters based on four alternative statistics (MedMedK, MedMeaK, MaxMedK, and MaxMeaK)[77] (Table 1). Based on the structure output, we separated the whole collection into five clusters (P1–P5). Winter-type Hungarian (European), spring-type Asian (Indian and Pakistani), and winter-type Turkish genotypes were dominant in clusters P1, P3, and P4, respectively. Cluster P2 contained mixed-type genotypes of different origins while cluster P5 was dominated by spring type NDSU advanced breeding lines of American origin and fiber type (Supplementary Table S1,

Structure run #	Burn-in lengths	MCMC lengths	Number of clusters (K)	Number of Reps	Number of clusters				
					ΔK[α]	Med MedK[β]	Med MeaK[β]	Max MedK[β]	Max MeaK[β]
1	5000	5000	10	10	9	4	4	5	5
2	10,000	10,000	10	10	9	4	4	5	5
3	10,000	50,000	10	10	3	4	4	5	5
4	20,000	20,000	10	10	9	4	4	5	5
5	20,000	50,000	10	10	6	4	4	4	4
6	50,000	50,000	10	10	7	4	4	5	5
7	50,000	100,000	10	10	6	4	4	5	5

Table 1. Number of clusters of the genotype collection based on Delta K approach[76] and four alternative statistics[77] using different combinations of burn-in lengths and Markov Chain Monte Carlo (MCMC) lengths. [α]The ad hoc ΔK method. [β] the median (MedMedK and MaxMedK) or mean (MedMeaK and MaxMeaK) estimators to determine the number of cluster (K).

Supplementary Fig. S5). Here, the type and origin of all genotypes were mentioned according to GRIN-Global database (https://www.grin-global.org/).

We also performed principal component analysis (PCA) to show the genetic similarity among genotypes. The first two axes explained 18% of the total observed variation. The output of principal component analysis was in line with that of structure analysis (Fig. 1A). A pairwise F_{st} comparison indicated genetic divergence among clusters. All combinations showed significant pairwise F_{st} comparison at $p < 0.01$. We found $F_{st} \geq 0.20$ for all combinations except combinations P1 and P2, P2 and P4. Cluster P3 showed maximum divergence ($F_{st} \geq 0.30$) from all other clusters (Table 2).

Moreover, Gaussian finite mixture model-based clustering of the whole collection using phenotypic data yielded four clusters (C1 to C4) (Fig. 1B). Cluster C1, C3, and C4 contained genotypes of different types and origins, while cluster C2 contained the highest number of genotypes dominated by spring type NDSU advanced breeding lines (Supplementary Table S1). Phenotype-based clustering was not consistent with genotype-based clustering (Fig. 1C, D) i.e., genotypic clusters containing genotypes belong to different phenotypic clusters and vice-versa.

In the whole collection, LD decayed to its half maximum within < 21 kb. LD decay rate varied according to chromosome (Supplementary Fig. S6), which was slowest in chromosomes Lu1 and Lu3 (32 kb) but was fastest in chromosomes Lu7 and Lu8 (15 kb).

The efficiency of different genomic prediction models

We determined the efficiency of 14 genomic prediction models in terms of computing time requirement and predictive ability (Figs. 2 and 3). For all traits, GBLUP required less time (< 60 min.) except the trait thousand seed

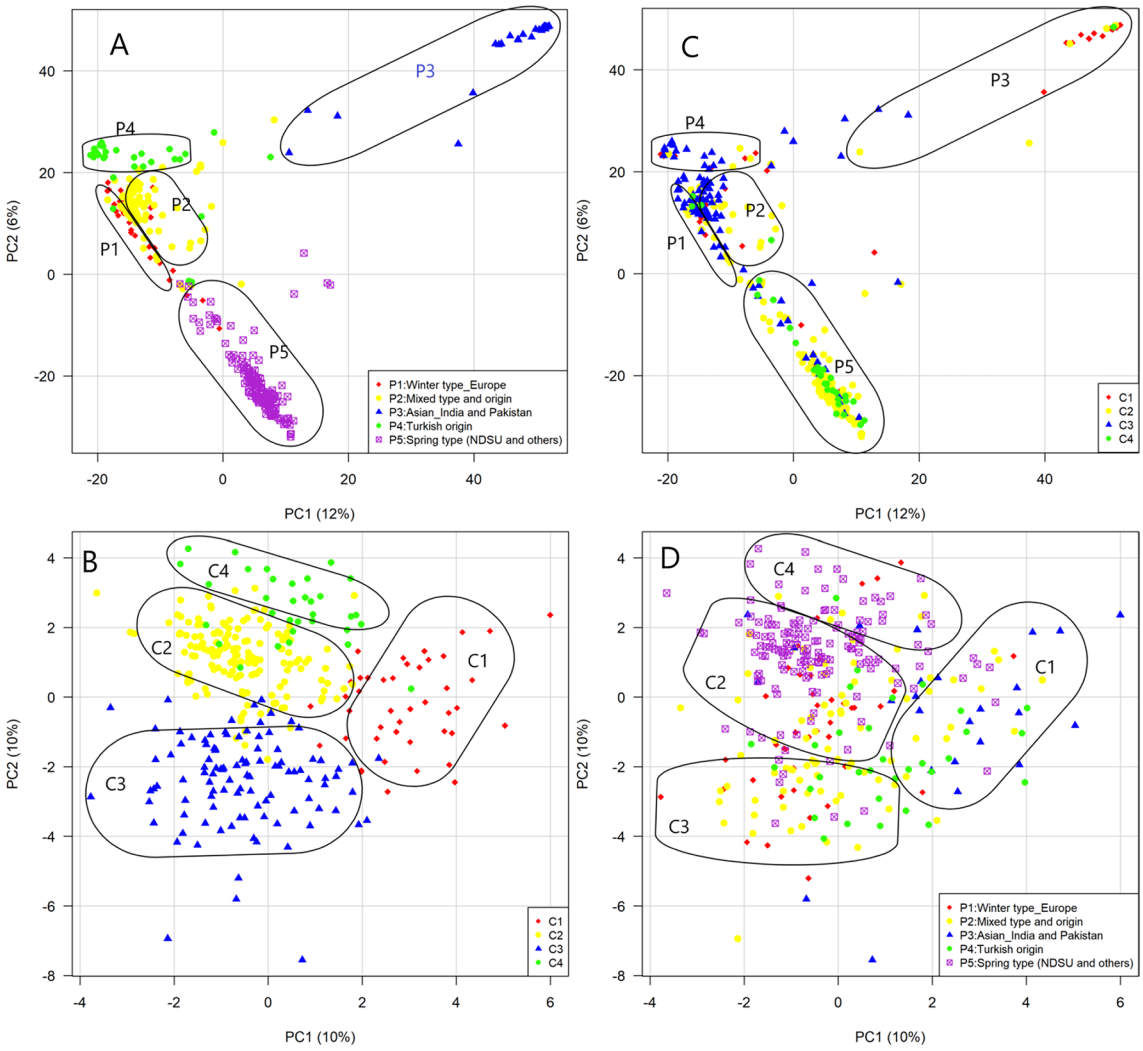

Figure 1. Genotype (SNP markers) and phenotype-based clustering of the whole collection. (**A**) Principal component analysis of SNP diversity based on genetic distance. Colors represent clusters identified at K = 5 in Supplementary Fig. S5. (**B**) Principal component analysis of the whole collection using phenotypic data. Colors represent groups identified by the Gaussian finite mixture model. (**C**) Genotypic clusters showing genotypes belong to different phenotypic groups. (**D**) Phenotypic groups showing genotypes belong to different genotypic clusters.

	Cluster pairwise F_{st}				
	P1	P2	P3	P4	P5
P1	0				
P2	0.13**	0			
P3	0.48**	0.38**	0		
P4	0.21**	0.13**	0.47**	0	
P5	0.24**	0.20**	0.50**	0.30**	0

Table 2. Genetic differentiation among different clusters. Diagonal values are pairwise F_{st} values based on 10,000 permutations using Arlequin v. 3.5. **indicates $p < 0.01$.

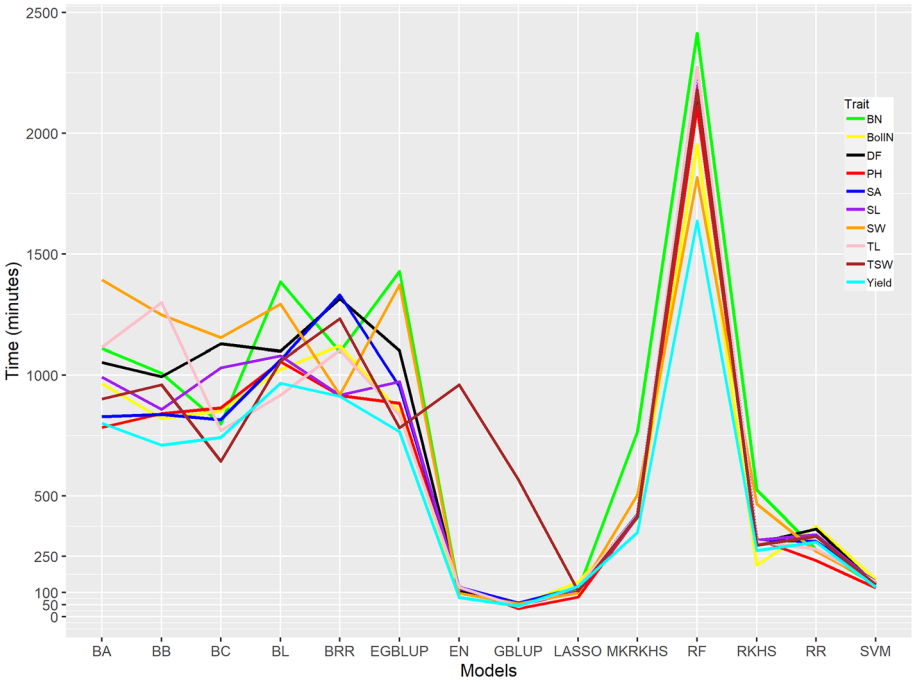

Figure 2. Line graph showing computing time required for running different models for different traits. Each model was replicated 100 times. DF is days to flowering, PH is plant height, TL is technical length, BN is branch number, BollN is boll number, TSW is thousand seed weight, SA is seed area, SW is seed width and SL is seed length.

weight (567 min.) and RF was the most time-demanding (1637–2414 min.) model. EGBLUP, a modification of the GBLUP model, which covers epistatic interactions, required more time (767–1428 min.) than GBLUP. Besides GBLUP, other less time-demanding models were EN (51–124 min.), LASSO (82–143 min), SVM (120–155 min.), and RR (233–374 min.). As time requirement by different models is heavily affected by computer configurations, we always used the same set-up for each model and the values presented here must be taken only for comparison.

The predictive ability values by different models vary significantly (p-value < 0.001) for each trait (Fig. 3, Supplementary Fig. S7). For all traits, SVM yielded the poorest predictive ability (− 0.03 to 0.09). LASSO and EN also resulted in low predictive ability for all traits. Model RR yielded the highest predictive ability for plant height (0.48), technical length (0.56), boll number (0.54), thousand seed weight (0.48), seed width (0.49), and seed yield (0.54). EGBLUP had the best predictive ability for days to flowering (0.19), and branch number (0.20), and RF had the best predictive ability for seed area (0.56), and seed length (0.56). The standard GBLUP model did not yield the highest predictive ability values for any traits and its magnitude was significantly (p-value < 0.001) lower compared to various models (Fig. 3 and Supplementary Fig. S7).

Predictive ability considering various marker subsets

We found significant ($p < 0.001$) variation among the predictive ability values for each trait according to various marker subsets (Table 3 and Supplementary Table S4). For all traits, predictive ability was lowest for 20 randomly selected markers, and it increased with the increment of marker numbers. Randomly selected 13,000 markers yielded the highest predictive ability for traits days to flowering, thousand seeds weight, seed length, and yield, whereas it was highest by 7000 randomly selected markers for traits plant height, technical length, and seed area.

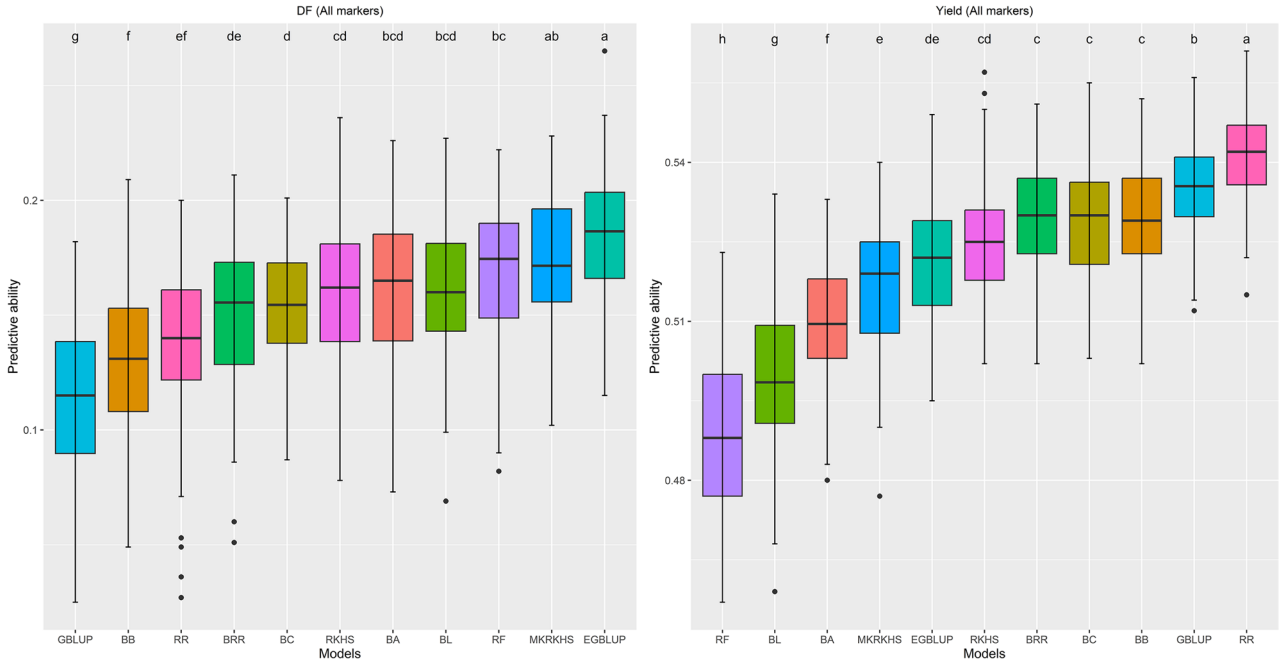

Figure 3. Boxplot showing the distribution of 100 replicates of predictive ability for days to flowering (DF) and seed yield using 14 models. SVM, LASSO, and EN model was not shown due to their low predictive ability. Boxplot for all traits were presented in Supplementary Fig. S7.

Marker selection criteria	No. of markers	Predictive ability									
		DF	PH	TL	BN	BollN	TSW	SA	SW	SL	Yield
All marker	26,171	0.19	0.48	0.56	0.20	0.54	0.48	0.56	0.49	0.56	0.54
All marker*	26,171	0.23	0.52	0.60	0.38	0.63	0.50	0.57	0.53	0.59	0.58
LD pruning	5362	0.15	0.48	0.56	0.21	0.54	0.48	0.55	0.49	0.54	0.54
RS	13,000	0.19	0.48	0.56	0.20	0.54	0.49	0.55	0.49	0.55	0.54
RS	7000	0.18	0.48	0.56	0.20	0.54	0.48	0.56	0.48	0.54	0.54
RS	3000	0.17	0.48	0.55	0.19	0.54	0.47	0.55	0.48	0.55	0.54
RS	1000	0.14	0.49	0.56	0.20	0.54	0.48	0.54	0.48	0.54	0.54
RS	200	0.14	0.45	0.54	0.15	0.51	0.45	0.52	0.46	0.52	0.50
RS	20	0.04	0.42	0.42	0.06	0.38	0.34	0.42	0.31	0.27	0.33

Table 3. Predictive ability based on randomly selected markers for different traits. In all cases, we assessed predictive ability using the best model identified in Fig. 3 and Supplementary Fig. S7. LD pruning = markers were selected based on chromosome-wise LD decay distance, RS = markers were randomly selected. DF is days to flowering, PH is plant height, TL is technical length, BN is branch number, BollN is boll number, TSW is thousand seed weight, SA is seed area, SW is seed width and SL is seed length. For each trait, predictive ability values by different marker subset varies significantly ($p < 0.001$), which was shown in detail in Supplementary Table S4. * indicate prediction accuracy calculated according to formula proposed by Ould Estaghvirou et al. (2013)[112].

Marker subset based on linkage disequilibrium decay showed the highest predictive ability for remaining traits. Utilization of the whole marker set did not yield the highest predictive ability values for any traits.

Marker subsets based on marker-trait associations (scenario-I to V) significantly (p- value < 0.001) affect the predictive ability values for all traits (Table 4 and Supplementary Table S5). Scenario- III yielded the lowest predictive ability for all traits, whereas it was highest by scenario-II for traits plant height (0.60), branch number (0.45), boll number (0.61), seed width (0.60), seed length (0.67) and yield (0.64), and by scenario- I for traits technical length (0.63), thousand seeds weight (0.72) and seed area (0.61). Scenario-IV yielded the highest (0.49) predictive ability for days to flowering at p-value ≤ 0.01. The predictive ability computed using all markers was better than that of scenario-III but was lower than that of scenario-I & II for all traits (Table 4). At all p-levels, the predictive ability for scenario-IV was better than that of scenario-V for days to flowering, technical length, branch number, and thousand seed weight, but opposite results were found for plant height, boll number, seed area, seed width, seed length, and seed yield. The marker number used for predictive ability computation varied according to traits and selection scenarios (Supplementary Table S3).

Marker selection criteria	*p*-value	Predictive ability									
		DF	PH	TL	BN	BollN	TSW	SA	SW	SL	Yield
All marker	–	0.19	0.48	0.56	0.20	0.54	0.48	0.56	0.49	0.56	0.54
Scenario- I	0.000103	0.28	0.58	0.63	0.28	0.58	0.72	0.61	0.52	0.67	0.61
Scenario-II*	–	0.47	0.60	0.61	0.45	0.61	0.64	0.51	0.60	0.67	0.64
Scenario-III*	–	0.02	0.27	0.17	0.14	0.39	0.33	0.41	0.33	0.40	0.27
Scenario-IV	0.001	0.45	0.49	0.55	0.45	0.53	0.51	0.46	0.52	0.47	0.53
Scenario-V	0.001	0.05	0.51	0.52	0.15	0.56	0.46	0.53	0.51	0.51	0.58
Scenario-IV	0.01	0.49	0.48	0.54	0.43	0.52	0.49	0.44	0.49	0.45	0.52
Scenario-V	0.01	0.11	0.51	0.53	0.11	0.56	0.46	0.53	0.51	0.51	0.58
Scenario-IV	0.05	0.48	0.47	0.53	0.44	0.51	0.48	0.43	0.48	0.43	0.51
Scenario-V	0.05	0.14	0.50	0.53	0.12	0.55	0.46	0.53	0.51	0.51	0.58

Table 4. Predictive ability based on markers, selected using marker-trait associations following scenario-I, II, III, IV, and V. Each scenario was discussed in detail in the method section. DF is days to flowering, PH is plant height, TL is technical length, BN is branch number, BollN is boll number, TSW is thousand seed weight, SA is seed area, SW is seed width and SL is seed length. In all cases, we assessed predictive ability using the best model identified in Fig. 3 and Supplementary Fig. S7. For each trait, predictive ability values by different marker subset varies significantly ($p < 0.001$), which was shown in detail in Supplementary Table S5. The number of markers used for different scenarios were mentioned in Supplementary Table S3. * In the case of scenario-II & III, the *p*-value was not mentioned as it varies according to traits.

Predictive ability considering population structure

We partitioned the whole collection into five (P1–P5) (Fig. 1A) and four (C1–C4) (Fig. 1B) clusters based on marker genotype and phenotypic data, respectively. To increase the level of relatedness among genotypes, genotypes belonging to clusters P3 and (P3 + P4) were discarded from the whole collection as these two clusters showed the greatest divergence from other clusters and incorporated Q-matrix in the model. In all cases, predictive ability was lower than that resulting from using the whole collection for all traits (Table 5). Predictive ability within each cluster was always lower than that of using the whole collection for all traits. The magnitude of genotype number within each cluster was not reflected by the magnitude of predictive ability i.e., clusters having more genotypes exhibited both high and low predictive ability for different traits and vice-versa (Table 5).

Stratified sampling and predictive ability assessment were done for all traits according to M-I and M-II. In the case of M-I, the predictive ability based on different sample sizes was better than that of the whole set (Fig. 4A). Sample size 125 yielded the best predictive ability for branch number, boll number, thousand seed

Clusters	Genotype number	Predictive ability									
		DF	PH	TL	BN	BollN	TSW	SA	SW	SL	Yield
Whole set (WS)	337	0.19	0.48	0.56	0.20	0.54	0.48	0.56	0.49	0.56	0.54
WS-P3[a]	312	0.19	0.42	0.52	0.17	0.52	0.49	0.55	0.46	0.55	0.52
WS-(P3 + P4)[b]	280	0.18	0.33	0.40	0.18	0.43	0.49	0.52	0.45	0.53	0.42
SP+	337	0.17	0.36	0.34	0.07	−0.10	0.40	0.41	0.36	0.45	−0.03
P1	41	0.01	−0.32	−0.37	0.33	−0.16	0.48	0.42	0.35	0.38	−0.15
P2	83	0.07	−0.01	−0.01	0.09	−0.04	0.04	0.08	0.06	0.09	0.17
P3	25	−0.27	−0.44	−0.47	−0.25	−0.44	0.34	0.49	0.46	0.30	−0.43
P4	32	0.07	−0.28	−0.36	−0.19	0.22	0.12	0.11	0.18	−0.02	0.0002
P5	156	0.22	0.11	0.14	0.11	0.32	0.22	0.13	0.09	0.18	0.19
C1	43	−0.29	−0.25	−0.09	0.10	−0.28	−0.23	0.10	−0.12	0.06	0.04
C2	156	0.18	−0.001	0.20	0.14	0.25	0.08	0.15	0.09	0.15	0.14
C3	105	0.20	0.22	0.28	0.13	0.25	0.31	0.19	0.06	0.07	0.21
C4	33	0.09	0.27	0.20	−0.29	−0.26	0.25	0.05	0.03	0.14	−0.25

Table 5. Comparison of predictive ability based on population structure. In all cases, we assessed predictive ability using the best model identified in Fig. 3 and Supplementary Fig. S7 using all markers. [a] Genotypes belonging to cluster P3 were discarded for analysis. [b] Genotypes belonging to clusters P3 and P4 were discarded for analysis. SP+ refers to the population structure addressed in the model. P denotes the cluster identified by structure analysis using SNP marker data in Fig. 1A. C denotes the cluster identified by the Gaussian finite mixture model based on phenotypic data in Fig. 1B. DF is days to flowering, PH is plant height, TL is technical length, BN is branch number, BollN is boll number, TSW is thousand seed weight, SA is seed area, SW is seed width and SL is seed length.

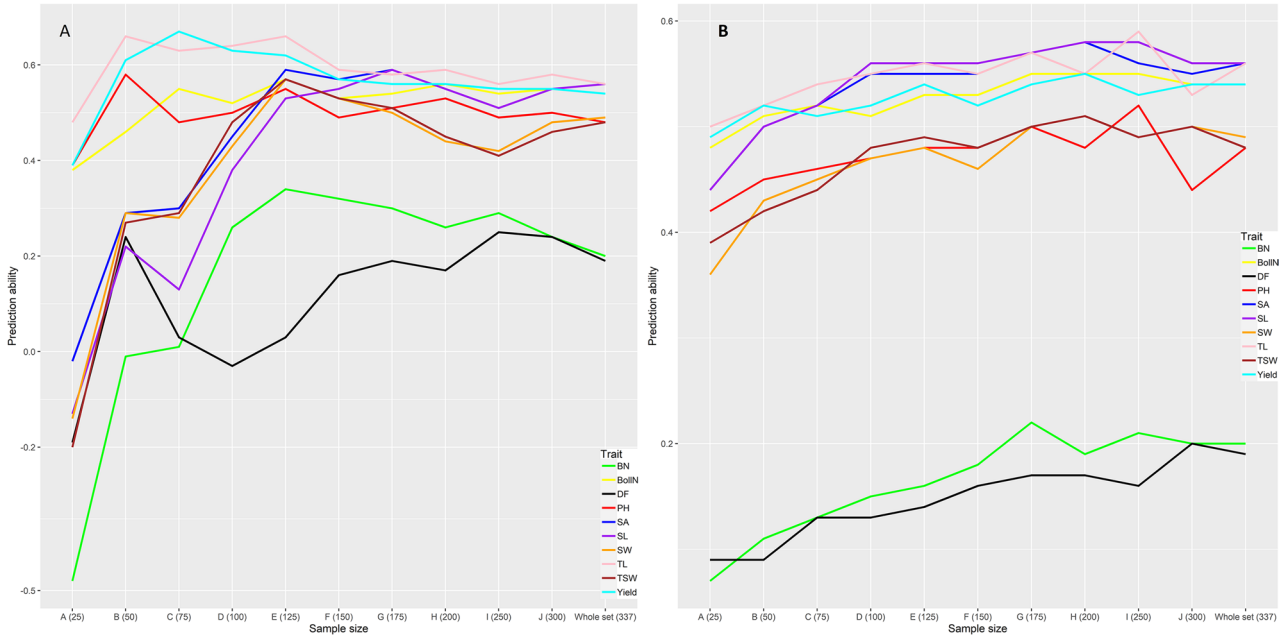

Figure 4. Stratified sampling effects on predictive ability. (**A**) Sampling and prediction ability were calculated according to M-I. Each sample was a training set, and the remaining genotype was a validation set. (**B**) Sampling and prediction ability were calculated according to M-II. Five-fold cross-validation was done within each sample. DF is days to flowering, PH is plant height, TL is technical length, BN is branch number, BollN is boll number, TSW is thousand seed weight, SA is seed area, SW is seed width and SL is seed length.

weight, seed width, and seed length. For days to flowering, plant height, and technical length, we found the best predictive ability at sample size 50, whereas it was better for seed length and seed yield at sample size 175 and 75, respectively. Likewise, M-I, as in the case of M-II, predictive ability was better based on different sample sizes than that for the whole set across traits (Fig. 4B). For days to flowering, plant height, technical length, branch number, and boll number the predictive ability reached a plateau at sample size 175, whereas, for thousand seed weight, seed area, seed width, seed length, and seed yield the same happened at sample size 200. In most cases, predictive ability based on M-I was better than that based on M-II.

Indirect predictive ability

The predictive ability of any target trait was calculated using the GEBVs of another trait (Table 6). Trait combinations with the best positive correlation resulted in better predictive ability, except for the combinations of days to flowering and plant height, days to flowering and branch number, and plant height and branch number. For instance, the predictive ability for seed yield based on GEBVs of plant height, technical length, and boll number

Traits	DF	PH	TL	BN	BollN	TSW	SA	SW	SL	Yield
DF	0.19	−0.16	−0.16	−0.09	−0.01	−0.05	−0.07	−0.08	−0.02	0.00
PH	0.36	0.48	0.55	0.11	0.43	−0.20	−0.27	−0.15	−0.30	0.46
TL	0.07	0.78	0.56	0.10	0.48	−0.21	−0.30	−0.17	−0.33	0.50
BN	0.46	0.69	0.37	0.20	0.15	0.21	0.21	0.26	0.17	0.15
BollN	−0.01	0.27	0.31	0.14	0.54	−0.19	−0.31	−0.18	−0.32	0.55
TSW	−0.04	−0.21	−0.21	−0.12	−0.04	0.48	0.55	0.48	0.55	−0.22
SA	−0.01	−0.18	− −0.21	−0.05	−0.17	0.3	0.56	0.49	0.57	−0.27
SW	−0.05	−0.16	−0.10	−0.10	−0.06	0.70	0.84	0.49	0.53	−0.20
SL	0.02	−0.24	−0.22	−0.12	−0.14	0.74	0.91	0.82	0.56	−0.29
Yield	−0.01	0.25	0.38	0.06	0.73	−0.07	−0.25	−0.12	−0.20	0.54

Table 6. Indirect predictive ability values for traits considering correlated traits. Diagonal values are direct predictive ability for traits. Above diagonal values are indirect predictive ability values for traits considering the GEBVs of correlated traits shown in column. Below diagonals are correlation coefficient values among traits. In all cases, we assessed predictive ability using the best model identified in Fig. 3 and Supplementary Fig. S7 using all markers. DF is days to flowering, PH is plant height, TL is technical length, BN is branch number, BollN is boll number, TSW is thousand seed weight, SA is seed area, SW is seed width and SL is seed length.

was 0.46, 0.50, and 0.55, respectively, as seed yield showed a better correlation to plant height (0.25), technical length (0.38), and boll number (0.73). We found similar results for other traits also.

Discussion

Phenotypic variability

Here, we investigated seed yield and nine other agronomic traits such as days to flowering, plant height, technical length, branch number, boll number, thousand seed weight, seed area, seed width, and seed length. These traits play an important role in flax development, adaptation, domestication, and improvement[113]. All these traits had continuous variation in all environments suggesting polygenic inheritance. Correlations among polygenic traits occur due to linkage and/or pleiotropic effect[114]. The better the correlations among traits the more likely it is that breeders can indirectly select one trait based on other traits with high heritability. In the current study, we found a very good positive correlation between plant height and technical length, and among seed-related traits. This finding will allow breeders to phenotype only plant height or technical length and any seed-related traits for further research using this germplasm collection, which will greatly reduce the phenotyping and analysis workload. Among the traits, boll number had the best correlation with seed yield. Previous studies[115,116] also confirmed the most direct contribution of boll or fruit number to flax seed yield. Days to flowering always had a negative correlation to seed yield and seed-related traits, which leads to the assumption that there is a possibility of exhausting more photosynthetic carbohydrates by late flowering flax genotypes for vegetative growth rather than seed formation and development. Seed traits such as thousand seed weight, seed area, seed width, and seed length were also negatively correlated to seed yield, which is consistent with previous studies[117–119].

Population structure

Structural variation and phenotypic diversity in a collection are inevitable when a breeder deals with a germplasm collection of different origins, types, and sources. Structural variation occurs due to allelic diversity present in the collection, whereas phenotypic variation is linked to this allelic diversity as well as environmental variations. Structure presence in a population influences its conservation and utilization and affects the output of genome-wide association analysis and genomic prediction. Population structure is influenced by mating strategy, mutation, selection, and gene flow[122]. The clear-cut separation of Asian (P3) and Turkish (P4) genotypes from others indicates that the geographic distance accelerates genetic differentiation by hindering gene flow. The presence of variable types of genotypes having mixed origin in sub-population P2 supports the hypothesis of active exchange of germplasm among European countries[123] as well as among other countries. The grouping of all NDSU-released varieties and advanced breeding lines and Canadian genotypes under the same sub-population P5 was due to shared ancestors and exchange of germplasm between the USA and Canada[124]. The fiber-type genotypes were in this sub-population as they were part of the parental set used for developing advanced breeding lines. We also partitioned the germplasm collection into four groups using ten quantitative agronomic traits. However, no clear-cut phenotypic clustering pattern according to types and origins was observed; the spring-type NDSU advanced breeding lines dominate one group, whereas spring-types of other origins and fiber type cluster together in another group. The pattern of the genotypic clusters was not reflected by phenotypic grouping and vice versa (Fig. 1). The mismatch between genotypic and phenotypic clustering output was also reported in flax[125], winged yam[126], and durum wheat[127]. This mismatch could be improved by incorporating more plant features i.e., traits (qualitative and quantitative) and diverse environments in further studies.

Genomic prediction

The studied germplasm collection showing considerable genetic diversity will resist genetic erosion, boost genetic gain, and serve as a source of valuable genes for further improvement. As the studied traits had continuous variations, relying on phenotypic evaluation alone can be expensive, laborious, and time-intensive. However, the low genotyping cost, relatively accurate genotyping, and efficient computational algorithms increase the opportunities to evaluate and utilize this collection for genomic prediction, which will reduce the time and cost associated with traits evaluation[128–132].

Comparing genomic prediction models

Various genomic prediction models are available, which can capture both linear (additive) and non-linear (epistasis and dominance) effects. In this study, we used predictive ability and time requirement to compare 14 different models. Breeders can also use prediction accuracy to compare the models, which can be calculated by dividing the predictive ability values by the square root of the corresponding traits' heritability[112]. Although prediction accuracy across all traits was better than corresponding predictive ability values (Table 3), we chose predictive ability as criteria to compare models, since there are possibilities of estimating and interpreting heritability poorly[133,134].

We found predictive ability values per trait by different models varied significantly, which was opposite[36,42] as well as similar[35,51,137,138] to previous reports. For example, Bari et al. (2021)[36] determined the predictive ability for six traits in a diverse pea germplasm collection using five different models. They observed almost the same predictive ability values across traits by different models. On the other hand, Azodi et al. (2019)[35], investigated 12 linear and non-linear models for different traits in six species and concluded that predictive ability by different models varies significantly for all traits. Yu et al. (2022)[51], Phumichai et al. (2022)[137], and Roorkiwal et al. (2016)[138] revealed the same phenomenon in rice, cassava, and chickpea, respectively. These reports and our findings confirmed that no single model worked best for all traits i.e., specific models are good for specific traits. Among all models, RR yielded the highest predictive ability for most of the traits, whereas it was lowest by SVM for all traits. A similar performance of SVM was found in wheat[34,135], but the opposite scenario in maize[136]. The differences in predictive ability values by models per trait and among traits were because of variation in the

underlying algorithm of models and the unique complex biology shaping the traits[129,139]. Gene action (additive, dominance, and epistasis) affects prediction accuracy for traits[141,142]. Empirical and theoretical evidence indicates that the lion share of genetic variance is additive, though gene action is not[143,144]. Momen et al. (2018)[37] reported that linear or parametric, and non-linear or non-parametric models outperform for traits under additive and non-additive gene action, respectively. In our study, for most of the traits model, RR resulted in the highest predictive ability values, which conferred that these traits were under additive gene action. Apart from this, linear model EGBLUP and non-linear model RF yielded the best predictive ability for days to flowering, branch number, seed area, and seed length. Epistatic gene action may shape these traits as outperforming models capture epistatic gene interaction[37,91]. In our study, the predictive ability values of different traits were proportionate to traits' heritability, indicating heritability affects genomic prediction, which was confirmed by previous reports in many crops[42,43,140]. In our study, branch number having low heritability showed poor predictive ability, whereas both were higher for other traits except days to flowering. In the case of days to flowering, though having high heritability, it showed low predictive ability. This may happen as there is a possibility of estimating heritability poorly[133,134]. Compared to our findings, Lan et al. (2020)[62] found better predictive ability for days to maturity, but lower predictive ability for seed yield in a bi-parental flax population of 260 lines. The lower predictive ability for seed yield was also supported by You et al. (2016) [63] in three different bi-parental flax populations. Overall findings indicate that the breeder should test various models for different traits to select the best-fitted model.

Marker density effect on predictive ability

Cost-effective next-generation sequencing techniques and the availability of high-quality reference genome have enabled breeders to extract informative genetic markers in prolific numbers. Utilizing these resources and high-performance computing facilities, breeders can feed the models with a huge number of markers. Although many models have been developed to handle the over-parameterization problem (marker number > > observation number) in genomic selection, previous reports confirmed that adding more markers after a certain number did not improve predictive ability. For example, in a wheat panel of 760 lines, predictive ability reached a plateau above around 5000 randomly selected markers[34], whereas in maize natural and bi-parental populations it requires 7000 and 2000 randomly selected markers, respectively[141]. We found the same trend of predictive ability in this study where a plateau was obtained around 1000–3000 randomly selected markers. In a recent simulation study, Chang et al. (2018)[44] achieved the same prediction accuracy by using 0.5 to 1% of all markers compared to that of the whole marker set (200,000). In our case, it happened at around 26% (7000) of total markers. This finding indicates that the predictive ability by marker subset capturing all QTL information and by the whole marker set will be the same. Our finding confirmed this, where predictive ability by a marker subset (5362) based on LD decay distance was higher than that by the whole marker set. In this study, more markers were required to obtain maximum predictive ability, though it required only 256 markers in a wheat bi-parental population[145], and 1000–1200 markers in a soybean varietal collection of 235 individuals[146]. This discrepancy in marker numbers among various research was due to the nature of the studied population and LD decay pattern. Although fewer markers can capture all QTL information in a bi-parental and varietal collection due to slow LD decay, more markers were required in this germplasm collection due to rapid LD decay.

Marker-trait association effect on predictive ability

In this study, the GWAS-derived significant SNP subset yielded better predictive ability values than that by the randomly selected marker subset and, surprisingly, even better predictive ability than that by the whole marker set across all traits. This finding was in line with results obtained by other authors in flax[62], wheat[34,45], maize[46,147], and spinach[148]. This overestimation only prevailed when marker selection was made on the whole population (training set + validation set), instead of the training set only, indicating a clear overfitting. That is why in practice, utilization of the marker set yielded by GWAS considering the whole collection is not recommended. This overfitting did not occur when marker selection was done based on a one-way ANOVA approach either using the whole collection or training set across all traits, except for days to flowering and branch number. This was due to capturing more markers (≥ 2000) by the one-way ANOVA approach, as more markers dissolve overfitting by acting like random selection. For future studies, breeders can use marker subsets based on LD decay distance rather than random selection and GWAS-based selection, which will ensure maximum predictive ability and will reduce run time.

Population structural effect on predictive ability

Because the availability of diverse genetic materials in a breeding program ensures its sustainability, breeders strive to improve predictive ability while maintaining genetic diversity. The studied genotypic collection has substantial genotypic and phenotypic diversity (Fig. 1). We corrected structural variation by discarding most divergent clusters and incorporating the population Q-matrix during analysis. In both cases, predictive ability did not improve across traits. Similar phenomena were found in wheat[149,150], pea[36,151], barley[152] and maize[47,153]. This finding confirms that for quantitative traits, reducing genetic structural variation increases genetic homogeneity in the collection, but not phenotypic homogeneity. We also observed very low to negative predictive ability across traits within each genotypic and phenotypic cluster, though Haile et al. (2021)[154] found moderate to high prediction accuracies within wheat subpopulations. The reason behind this was the inconsistency between genotypic and phenotypic variance i.e., genotypes of specific genetic clusters were grouped under different phenotypic clusters and the same happened to the genotypes of specific phenotypic clusters (Fig. 1C, D). Smaller population sizes in some clusters may also contribute to this.

In this study, the reduction of predictive ability due to the correction of population structure has driven us to follow a stratified sampling approach, as previous research indicates that stratified sampling[48] or composite

sampling[155] increases predictive ability. The same has happened in our research as stratified sampling (M-I and M-II) yielded better predictive ability compared to the whole collection. For all traits, in the case of M-I, a small sample size (50–125) yielded better predictive ability compared to the whole collection, whereas it gradually increased with the increment of training size in the case of M-II. In terms of predictive ability, M-I was more productive than M-II, which confirmed that small-sized stratified sample is strong enough to capture diversity as well as maintain good predictive ability in a diverse germplasm collection. In addition, breeders can use genotypes as parents from a particular sample yielding the higher predictive ability to make Multi-parent Advanced Generation Intercrosses (MAGIC) populations for future studies because the relatedness among individuals of the training set and target set accelerates predictive ability[38,39,156].

Indirect genomic prediction

In this study, traits had both positive and negative correlations with each other. Breeders can utilize information on correlated traits to predict the target trait using a multi-trait genomic prediction approach. Many previous studies exhibited benefits[40,41,157,158] and no benefit[159–161] of multi-trait over single-trait prediction approach in different crops. However, one of the major limitations of the multi-trait approach is that breeders need to phenotype multiple correlated traits, which is expensive and laborious. To overcome this limitation, breeders can use an indirect genomic prediction approach i.e., prediction of genotype for target trait by using a correlated single trait. The benefit of the indirect approach accelerates if the correlated trait possesses high heritability and is easy to phenotype at the crop's early stages of development and vice-versa for the focal trait. Our findings revealed the benefit of indirect genomic selection as indirect predictive ability based on highly correlated traits was very close to its single-trait predictive ability. A similar result was found by Fernandes et al. (2018)[161] in sorghum where they reported the indirect prediction accuracy for biomass yield by plant height was similar to its single-trait and multi-trait prediction accuracies. Our findings will help breeders to reduce workload by performing indirect selection for expensive or labor-intensive focal traits by phenotyping early expressed correlated simple-to-measure traits at an early stage of the breeding pipeline.

Conclusion

In this study, a rigorous investigation of various key factors affecting genomic predictive ability was conducted by comparing fourteen different models. The results indicated that models have a significant effect on predictive ability and no single model worked best across all traits, though model RR shows the potentiality by yielding higher predictive ability values for most of the trait. It is better to compare various models to choose the best one for any trait. Predictive ability reaches a plateau around a certain marker density and shows similarity with the whole marker set when choosing markers covering all QTL information. In the diverse flax collection used for this study, the small sample size representing population structure was strong enough to boost predictive ability compared to the whole collection across all traits. Along with this, indirect selection for seed yield considering correlated traits also holds potential for applied breeding efforts. This research presented herein will equip the plant breeders to efficiently design various aspects of genomic prediction to gain increased selection accuracy, which will subsequently accelerate the program's genetic gain.

Data availability

All raw sequence data and variant data are available in the NCBI and EVA repositories. The accession IDs for them are PRJNA979944 (https://www.ncbi.nlm.nih.gov/sra/PRJNA979944) and PRJEB62432 (https://www.ebi.ac.uk/eva/?eva-study=PRJEB62432), respectively. The Phenotypic datasets and R scripts used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 28 July 2023; Accepted: 31 January 2024
Published online: 08 February 2024

References

1. Muir, A. D. & Westcott, N. D. *Flax: The Genus Linum* (CRC Press, 2003).
2. Touré, A. & Xueming, X. Flaxseed Lignans: Source, biosynthesis, metabolism, antioxidant activity, bio-active components, and health benefits. *Compr. Rev. Food Sci. Food Saf.* **9**, 261–269. https://doi.org/10.1111/J.1541-4337.2009.00105.X (2010).
3. Westcott, N. D. & Muir, A. D. Flax seed lignan in disease prevention and health promotion. *Phytochem. Rev.* **2**, 401–417 (2003).
4. Przybylski, R. Flax oil and high linolenic oils. *Bailey's Ind. Oil Fat Prod.* **2**, 281–301 (2005).
5. Morris, D. H. *The Novel Egg: Opportunities for Flax in Omega-3 Egg Production* (Flax Council of Canada, 2003).
6. Ndou, S. P., Kiarie, E., Walsh, M. C. & Nyachoti, C. M. Nutritive value of flaxseed meal fed to growing pigs. *Anim. Feed. Sci. Technol.* **238**, 123–129 (2018).
7. Cullis, C. *Wild Crop Relatives: Genomic and Breeding Resources* 177–189 (Springer, 2011).
8. FAOSTAT. Food and agriculture organization of the United Nations. Statistical database (2022).
9. USDA-NASS. United States Census of Agriculture (2022).
10. Berglund, D. R. & Zollinger, R. K. Flax production in North Dakota. North Dakota Agricultural Experimental Station, Extension Service North Dakota, Fargo, North Dakota, USA. Bull. A-1038. 12 p. (2002).
11. Duguid, S., Lafond, G., McAndrew, D. W., Rashid, K. Y. & Ulrich, A. *Growing Flax: Production, Management & Diagnostic Guide* (Flax Council of Canada, 2007).
12. Xu, Y. & Crouch, J. H. Marker-assisted selection in plant breeding: From publications to practice. *Crop. Sci.* **48**, 391–407 (2008).
13. Nihad, S. A. I. *et al.* Linkage of SSR markers with rice blast resistance and development of partial resistant advanced lines of rice (*Oryza sativa*) through marker-assisted selection. *Physiol. Mol. Biol. Plants* https://doi.org/10.1007/S12298-022-01141-3 (2022).
14. Sun, L. *et al.* Robust identification of low-Cd rice varieties by boosting the genotypic effect of grain Cd accumulation in combination with marker-assisted selection. *J. Hazard Mater.* **424**, 127703 (2022).
15. Alsaleh, A. *et al.* Marker-assisted selection and validation of DNA markers associated with cadmium content in durum wheat germplasm. *Crop Pasture Sci.* https://doi.org/10.1071/CP21484 (2022).

16. Soriano, M. *et al.* Identification and characterisation of stripe rust resistance genes Yr66 and Yr67 in wheat cultivar VL Gehun 892. *Agronomy* **12**, 318 (2022).
17. Yadav, P. S. *et al.* Enhanced resistance in wheat against stem rust achieved by marker assisted backcrossing involving three independent Sr genes. *Curr. Plant Biol.* **2**, 25–33 (2015).
18. Yang, R., Yan, Z., Wang, Q., Li, X. & Feng, F. Marker-assisted backcrossing of *lcyE* for enhancement of *proA* in sweet corn. *Euphytica* **214**, 1–12 (2018).
19. Hao, X., Li, X., Yang, X. & Li, J. Transferring a major QTL for oil content using marker-assisted backcrossing into an elite hybrid to increase the oil content in maize. *Mol. Breed.* **34**, 739–748 (2014).
20. Yathish, K. R. *et al.* Introgression of the low phytic acid locus (lpa2) into elite maize (*Zea Mays* L.) inbreds through marker-assisted backcross breeding (MABB). *Euphytica* **218**, 127. https://doi.org/10.21203/rs.3.rs-1293507/v1 (2022).
21. Becker, H. C. & Bernardo, R. A model for marker-assisted selection among single crosses with multiple genetic markers. *Theor. Appl. Genet.* **97**, 473–478 (1998).
22. Bernardo, R. *Breeding for Quantitative tRaits in Plants* 3rd edn. (Stemma Press, 2020).
23. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
24. Lorenz, A. J. *et al.* Genomic selection in plant breeding: Knowledge and prospects. *Adv. Agron.* **110**, 77–123 (2011).
25. Schaeffer, L. R. Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* **123**, 218–223 (2006).
26. Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **4**, 250–255 (2011).
27. Long, N., Gianola, D., Rosa, G. J. M. & Weigel, K. A. Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* **123**, 1065–1074 (2011).
28. de Los Campos, G. *et al.* Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**, 375–385 (2009).
29. Park, T. & Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **103**, 681–686. https://doi.org/10.1198/016214508000000337 (2012).
30. Crossa, J. *et al.* Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713–724 (2010).
31. Ober, U. *et al.* Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics* **188**, 695–708 (2011).
32. Long, N., Gianola, D., Rosa, G. J. M. & Weigel, K. A. Marker-assisted prediction of non-additive genetic values. *Genetica* **139**, 843–854 (2011).
33. de Los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A. & Crossa, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* **92**, 295–308 (2010).
34. Charmet, G., Tran, L. G., Auzanneau, J., Rincent, R. & Bouchet, S. BWGS: A R package for genomic selection and its application to a wheat breeding programme. *PLoS One* **15**, e0222733 (2020).
35. Azodi, C. B. *et al.* Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 Genes Genomes Genet.* **9**, 3691–3702 (2019).
36. Bari, M. A. A. *et al.* Harnessing genetic diversity in the USDA pea germplasm collection through genomic prediction. *Front. Genet.* **12**, 2273 (2021).
37. Momen, M. *et al.* Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci. Rep.* **8**, 1–11 (2018).
38. Riedelsheimer, C., Brotman, Y., Méret, M., Melchinger, A. E. & Willmitzer, L. The maize leaf lipidome shows multilevel genetic control and high predictive value for agronomic traits. *Sci. Rep.* **3**, 1–7 (2013).
39. Rutkoski, J. *et al.* Efficient use of historical data for genomic selection: A case study of stem rust resistance in wheat. *Plant Genome* **8**, 1. https://doi.org/10.3835/plantgenome2014.09.0046 (2015).
40. Bhatta, M. *et al.* Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (*Hordeum vulgare* L.). *G3 Genes Genomes Genet.* **10**, 1113–1124 (2020).
41. Velazco, J. G. *et al.* Genomic prediction of grain yield and drought-adaptation capacity in sorghum is enhanced by multi-trait analysis. *Front. Plant Sci.* **10**, 997 (2019).
42. Spindel, J. *et al.* Genomic selection and association mapping in rice (*Oryza sativa*): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite. Tropical rice breeding lines. *PLoS Genet.* **11**, e1004982 (2015).
43. Zhang, A. *et al.* Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Front. Plant Sci.* **8**, 1916 (2017).
44. Chang, L. Y., Toghiani, S., Ling, A., Aggrey, S. E. & Rekaya, R. High density marker panels, SNPs prioritizing and accuracy of genomic selection. *BMC Genet.* **19**, 1–10 (2018).
45. Sehgal, D. *et al.* Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Front. Plant Sci.* **11**, 197 (2020).
46. Rice, B. & Lipka, A. E. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *Plant Genome* **12**, 180052 (2019).
47. Guo, Z. *et al.* The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet.* **127**, 749–762 (2014).
48. Isidro, J. *et al.* Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* **128**, 145–158 (2015).
49. Frouin, J., Labeyrie, A., Boisnard, A., Sacchi, G. A. & Ahmadi, N. Genomic prediction offers the most effective marker assisted breeding approach for ability to prevent arsenic accumulation in rice grains. *PLoS One* **14**, e0217516 (2019).
50. Monteverde, E. *et al.* Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa* L.) grown in subtropical areas. *G3 Genes Genomes Genet.* **9**, 1519–1531 (2019).
51. Yu, P. *et al.* Genome-wide association study and genomic prediction for yield and grain quality traits of hybrid rice. *Mol. Breed.* **42**, 1–12 (2022).
52. Huang, M. *et al.* Use of genomic selection in breeding rice (*Oryza sativa* L.) for resistance to rice blast (*Magnaporthe oryzae*). *Mol. Breed.* **39**, 1–16 (2019).
53. Ben-Sadoun, S. *et al.* Economical optimization of a breeding scheme by selective phenotyping of the calibration set in a multi-trait context: Application to bread making quality. *Theor. Appl. Genet.* **133**, 2197–2212 (2020).
54. Merrick, L. F., Herr, A. W., Sandhu, K. S., Lozada, D. N. & Carter, A. H. Utilizing genomic selection for wheat population development and improvement. *Agronomy* **12**, 522 (2022).
55. Song, J. *et al.* Practical application of genomic selection in a doubled-haploid winter wheat breeding program. *Mol. Breed.* **37**, 1–15 (2017).
56. Hu, X. *et al.* Effectiveness of genomic selection by response to selection for winter wheat variety improvement. *Plant Genome* **12**, 180090 (2019).
57. Robert, P. *et al.* Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. *Theor. Appl. Genet.* **135**, 895–914 (2022).

58. Cerrudo, D. *et al.* Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front. Plant Sci.* **9**, 366 (2018).
59. Zhang, X. *et al.* Rapid cycling genomic selection in a multiparental tropical maize population. *G3 Genes Genomes Genet.* **7**, 2315–2326 (2017).
60. Fristche-Neto, R., Akdemir, D. & Jannink, J. L. Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theor. Appl. Genet.* **131**, 1153–1162 (2018).
61. Mastrodomenico, A. T., Bohn, M. O., Lipka, A. E. & Below, F. E. Genomic selection using maize ex-plant variety protection germplasm for the prediction of nitrogen-use traits. *Crop. Sci.* **59**, 212–220 (2019).
62. Lan, S. *et al.* Genomic prediction accuracy of seven breeding selection traits improved by QTL identification in flax. *Int. J. Mol. Sci.* **21**, 1577 (2020).
63. You, F. M., Booker, H. M., Duguid, S. D., Jia, G. & Cloutier, S. Accuracy of genomic selection in biparental populations of flax (*Linum usitatissimum* L.). *Crop. J.* **4**, 290–303 (2016).
64. He, L. *et al.* Evaluation of genomic prediction for pasmo resistance in flax. *Int. J. Mol. Sci.* **20**, 359 (2019).
65. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379 (2011).
66. Glaubitz, J. C. *et al.* TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**, e90346 (2014).
67. You, F. M. *et al.* Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. *Plant J.* **95**, 371–384 (2018).
68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
69. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
70. Federer, W. T. & Crossa, J. I.4 screening experimental designs for quantitative trait loci, association mapping, genotype-by environment interaction, and other investigations. *Front. Physiol.* **3**, 156 (2012).
71. Nôžková, J. et al. Descriptor list for flax–*Linum usitatissimum* L. Nitra: SPU (2011).
72. Cullis, B. R., Smith, A. B. & Coombes, N. E. On the design of early generation variety trials with correlated data. *J. Agric. Biol. Environ. Stat.* **11**, 381–393 (2006).
73. Covarrubias-Pazaran, G. Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* **11**, e0156744 (2016).
74. Taiyun, W. M. et al. Package 'corrplot' Title Visualization of a Correlation Matrix. (2017).
75. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
76. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).
77. Puechmaille, S. J. The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Mol. Ecol. Resour.* **16**, 608–627 (2016).
78. Li, Y. L. & Liu, J. X. StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour.* **18**, 176–177 (2018).
79. Jakobsson, M. & Rosenberg, N. A. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
80. Bradbury, P. J. *et al.* TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
81. Jakobsson, M., Edge, M. D. & Rosenberg, N. A. The relationship between FST and the frequency of the most frequent allele. *Genetics* https://doi.org/10.1534/genetics.112.144758 (2013).
82. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
83. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *R. J.* **8**, 289 (2016).
84. Villanueva, R. A. M. & Chen, Z. J. *ggplot: Elegant Graphics for Data Analysis* 2nd edn, 160–167 (Taylor & Francis, 2019). https://doi.org/10.1080/15366367.2019.156525417.
85. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2018).
86. Hill, W. G. & Weir, B. S. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33**, 54–78 (1988).
87. Desta, Z. A. & Ortiz, R. Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* **19**, 592–601 (2014).
88. Piepho, H. P. Ridge regression and extensions for genome wide selection in maize. *Crop. Sci.* **49**, 1165–1176 (2009).
89. Habier, D., Fernando, R. L. & Garrick, D. J. Genomic BLUP decoded: A look into the black box of genomic prediction. *Genetics* **194**, 597–607 (2013).
90. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
91. Jiang, Y. & Reif, J. C. Modeling epistasis in genomic selection. *Genetics* **201**, 759–768 (2015).
92. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
93. Usai, M. G., Goddard, M. E. & Hayes, B. J. LASSO with cross-validation for genomic selection. *Genet. Res. (Camb.)* **91**, 427–436 (2009).
94. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**(2), 301–320 (2005).
95. de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. & Calus, M. P. L. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345 (2013).
96. Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. Extension of the bayesian alphabet for genomic selection. *BMC Bioinform.* **12**, 1–12 (2011).
97. Pérez, P. & de Los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**, 483–495 (2014).
98. Gianola, D. & van Kaam, J. B. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**, 2289–2303 (2008).
99. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
100. Maenhout, S., de Baets, B., Haesaert, G. & van Bockstaele, E. Support vector machine regression for the prediction of maize hybrid performance. *Theor. Appl. Genet.* **115**, 1003–1013 (2007).
101. González-Recio, O., Rosa, G. J. M. & Gianola, D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest Sci.* **166**, 217–231 (2014).
102. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

103. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
104. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
105. Li, M. *et al.* Enrichment of statistical power for genome-wide association studies. *BMC Biol.* **12**, 1–10 (2014).
106. Wang, Q., Tian, F., Pan, Y., Buckler, E. S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS One* **9**, e107684 (2014).
107. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**, 825–830 (2012).
108. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**, e1005767 (2016).
109. Huang, M., Liu, X., Zhou, Y., Summers, R. M. & Zhang, Z. BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience* **8**, 1–12 (2019).
110. Wang, J. & Zhang, Z. GAPIT Version 3: Boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinform.* **19**, 629–640 (2021).
111. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **2005**(95), 221–227 (2005).
112. Ould, E. S. B. *et al.* Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics* **14**, 1–21 (2013).
113. Zhang, J. *et al.* Genomic comparison and population diversity analysis provide insights into the domestication and improvement of flax. *iScience* **23**, 100967 (2020).
114. Saltz, J. B., Hessel, F. C. & Kelly, M. W. Trait correlations in the genomics era. *Trends Ecol. Evol.* **32**, 279–290 (2017).
115. Çopur, O. & Demirel, U. Determination of correlation and path analysis among yield components and seed yield in oil flax varieties (*Linum usitatissimum* L.). *J. Biol. Sci.* https://doi.org/10.3923/jbs.2006.738.743 (2006).
116. Bibi, T., Mahmood, T., Mirza, Y. & Mahmood, T. Correlation studies of some yield related traits in linseed (*Linum usitatissimum* L.). *J. Agric. Res.* **51**, 121–132 (2013).
117. Worku, N., Heslop-Harrison, J. S. & Adugna, W. Diversity in 198 Ethiopian linseed (*Linum usitatissimum*) accessions based on morphological characterization and seed oil characteristics. *Genet. Resour. Crop. Evol.* **62**, 1037–1053 (2015).
118. Soto-Cerda, B. J. *et al.* Genomic regions underlying agronomic traits in linseed (*Linum usitatissimum* L.) as revealed by association mapping. *J. Integr. Plant Biol.* **56**, 75–87 (2014).
119. Soto-Cerda, B. J. *et al.* Assessing the agronomic potential of linseed genotypes by multivariate analyses and association mapping of agronomic traits. *Euphytica* **196**, 35–49 (2014).
120. Yu, Z., Fredua-Agyeman, R., Hwang, S.-F. & Strelkov, S. E. Molecular genetic diversity and population structure analyses of rutabaga accessions from Nordic countries as revealed by single nucleotide polymorphism markers. *BMC Genomics* **22**, 1–13 (2021).
121. Rahman, M., Hoque, A. & Roy, J. Linkage disequilibrium and population structure in a core collection of *Brassica napus* (L.). *PLoS ONE* **17**(3), e0250310. https://doi.org/10.1371/journal.pone.0250310 (2022).
122. Schaal, B. A., Hayworth, D. A., Olsen, K. M., Rauscher, J. T. & Smith, W. A. Phylogeographic studies in plants: Problems and prospects. *Mol. Ecol.* **7**, 465–474 (1998).
123. Maggioni, L. Flax genetic resources in Europe: Ad Hoc Meeting, 7–8 December 2001, Prague, Czech Republic. (Bioversity International, 2002).
124. Fu, Y.-B., Rowland, G. G., Duguid, S. D. & Richards, K. W. RAPD analysis of 54 North American flax cultivars. *Crop. Sci.* **43**, 1510–1515 (2003).
125. Choudhary, S. B. *et al.* Genetic diversity spectrum and marker trait association for agronomic traits in global accessions of *Linum usitatissimum* L. *Ind. Crops Prod.* **108**, 604–615 (2017).
126. Agre, P. *et al.* Phenotypic and molecular assessment of genetic structure and diversity in a panel of winged yam (*Dioscorea alata*) clones and cultivars. *Sci. Rep.* **9**, 1–11 (2019).
127. Royo, C. *et al.* Understanding the relationships between genetic and phenotypic structures of a collection of elite durum wheat accessions. *Field Crops Res.* **119**, 91–105 (2010).
128. Mascher, M. *et al.* Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nat. Genet.* **51**, 1076–1081 (2019).
129. Yu, X. *et al.* Genomic prediction of maize microphenotypes provides insights for optimizing selection and mining diversity. *Plant Biotechnol. J.* **18**, 2456–2465 (2020).
130. Yu, X. *et al.* Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat. Plants* **2**, 1–7 (2016).
131. Li, H., Rasheed, A., Hickey, L. T. & He, Z. Fast-forwarding genetic gain. *Trends Plant Sci.* **23**, 184–186 (2018).
132. Crossa, J. *et al.* Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **22**, 961–975 (2017).
133. Piepho, H. P. & Möhring, J. Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**, 1881–1888 (2007).
134. Dudley, J. W. & Moll, R. H. Interpretation and use of estimates of heritability and genetic variances in plant breeding1. *Crop Sci* **9**, 257–262 (1969).
135. Ornella, L. et al. Genomic prediction of genetic values for resistance to wheat rusts. *Plant Genome* 5, (2012).
136. Zhao, W. *et al.* Applications of support vector machine in genomic prediction in pig and maize populations. *Front. Genet.* **11**, 1537 (2020).
137. Phumichai, C. *et al.* Genome-wide association mapping and genomic prediction of yield-related traits and starch pasting properties in cassava. *Theor. Appl. Genet.* **135**, 145–171 (2022).
138. Roorkiwal, M. *et al.* Genome-enabled prediction models for yield related traits in chickpea. *Front. Plant Sci.* **7**, 1666 (2016).
139. Valluru, R. *et al.* Deleterious mutation burden and its association with complex traits in sorghum (*Sorghum bicolor*). *Genetics* **211**, 1075–1087 (2019).
140. Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb.)* **91**, 47–60 (2009).
141. Liu, X. *et al.* Factors affecting genomic selection revealed by empirical evidence in maize. *Crop. J.* **6**, 341–352 (2018).
142. Raffo, M. A. *et al.* Improvement of genomic prediction in advanced wheat breeding lines by including additive-by-additive epistasis. *Theor. Appl. Genet.* **135**, 965–978 (2022).
143. Mäki-Tanila, A. & Hill, W. G. Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**, 355–367 (2014).
144. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
145. Heffner, E. L. *et al.* Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* **4**, 65–75 (2011).
146. Ma, Y. *et al.* Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). *Mol. Breed.* **36**, 1–10 (2016).

147. Bian, Y. & Holland, J. B. Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. *Heredity* **118**, 585–593 (2017).
148. Shi, A. *et al.* Genome-wide association study and genomic prediction of white rust resistance in USDA GRIN spinach germplasm. *Hortic. Res.* **9**, uhac069 (2022).
149. Crossa, J. *et al.* Genomic prediction of gene bank wheat landraces. *G3 Genes Genomes Genet.* **6**, 1819–1834 (2016).
150. Norman, A., Taylor, J., Edwards, J. & Kuchel, H. Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3 Genes Genomes Genet.* **8**, 2889–2899 (2018).
151. Burstin, J. *et al.* Genetic diversity and trait genomic prediction in a pea diversity panel. *BMC Genomics* **16**, 1–17 (2015).
152. Thorwarth, P. *et al.* Genomic prediction ability for yield-related traits in German winter barley elite material. *Theor. Appl. Genet.* **130**, 1669–1683 (2017).
153. Lyra, D. H. *et al.* Controlling population structure in the genomic prediction of tropical maize hybrids. *Mol. Breed.* **38**, 1–17 (2018).
154. Haile, T. A. *et al.* Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. *Theor. Appl. Genet.* **134**, 381–398 (2021).
155. He, S. *et al.* Genomic prediction using composite training sets is an effective method for exploiting germplasm conserved in rice gene banks. *Crop. J.* https://doi.org/10.1016/J.CJ.2021.11.011 (2022).
156. Lorenz, A. & Smith, K. P. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop. Sci.* **55**, 2657–2667 (2015).
157. Schulthess, A. W., Zhao, Y., Longin, C. F. H. & Reif, J. C. Advantages and limitations of multiple-trait genomic prediction for Fusarium head blight severity in hybrid wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **131**, 685–701 (2018).
158. Lyra, D. H. *et al.* Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. *Mol. Breed.* **37**, 1–14 (2017).
159. dos Santos, J. P. R., de Castro Vasconcellos, R. C., Pires, L. P. M., Balestre, M. & von Pinho, R. G. Inclusion of dominance effects in the multivariate GBLUP model. *PLoS One* **11**, e0152045 (2016).
160. Schulthess, A. W. *et al.* Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theor. Appl. Genet.* **129**, 273–287 (2016).
161. Fernandes, S. B., Dias, K. O. G., Ferreira, D. F. & Brown, P. J. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* **131**, 747–755 (2018).

Acknowledgements

We thank Mr. Kreg Kercher (Department of Plant Sciences, NDSU, Fargo, ND) for helping in field planting and data collection. We also thank Justin D Faris (Research Geneticist, USDA-ARS, Fargo, ND) for providing the MARVIN seed analyzer. Technical assistance during data analyses from Md. Abdullah Al Bari (Department of Plant Sciences, NDSU, Fargo, ND), Jason D. Fiedler (Research Plant Molecular Geneticist, USDA-ARS, Fargo, ND), and Brant Bigger (Plant Physiologist, USDA-ARS, Fargo, ND) are gratefully acknowledged. This work used resources of the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University, which were made possible in part by NSF MRI Award No. 2019077.

Author contributions

A.H. and M.R. conceived and designed the study. A.H. conducted the data collection, curation, and analyses. A.H. did the interpretation of results with the help of M.R. and J.V.A. A.H. wrote the manuscript. M.R. is the principal investigator of the project. All authors participated in revising and editing the manuscript and approved the final version of the manuscript.

Funding

The study was funded by the U.S. Department of Agriculture—National Institute of Food and Agriculture (Hatch Project No. ND01581). The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-53462-w.

Correspondence and requests for materials should be addressed to M.R.