Multi-environment analysis enhances genomic prediction accuracy of agronomic traits in sesame

Idan Sabag[1,2], Ye Bi[2], Zvi Peleg[1,*], and Gota Morota[2,*]

[1]The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, Rehovot 7610001, Israel
[2]School of Animal Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Running title: Genomic prediction for sesame

Keywords: genomic prediction, Mediterranean climate, multi-environment, oilseed crop, sesame

* Corresponding author
E-mail: morota@vt.edu and zvi.peleg@mail.huji.ac.il

ORCID: 0000-0003-2560-6845 (IS), 0000-0001-7871-5856 (YB), 0000-0001-8063-1619 (ZP), and 0000-0002-3567-6911 (GM)

Email addresses: idan.sabag@mail.huji.ac.il (IS), yebi@vt.edu (YB), zvi.peleg@mail.huji.ac.il (ZP), and morota@vt.edu (GM)

Abstract

Sesame is an ancient oilseed crop containing many valuable nutritional components. Recently, the demand for sesame seeds and their products has increased worldwide, making it necessary to enhance the development of high-yielding cultivars. One approach to enhance genetic gain in breeding programs is genomic selection. However, studies on genomic selection and genomic prediction in sesame are limited. In this study, we performed genomic prediction for agronomic traits using the phenotypes and genotypes of a sesame diversity panel grown under Mediterranean climatic conditions over two growing seasons. We aimed to assess the accuracy of prediction for nine important agronomic traits in sesame using single- and multi-environment analyses. In single-environment analysis, genomic best linear unbiased prediction, BayesB, BayesC, and reproducing kernel Hilbert spaces models showed no substantial differences. The average prediction accuracy of the nine traits across these models ranged from 0.39–0.79 for both growing seasons. In the multi-environment analysis, the marker-by-environment interaction model, which decomposed the marker effects into components shared across environments and environment-specific deviations, improved the prediction accuracies for all traits by 15%–58% compared to the single-environment model, particularly when borrowing information from other environments was made possible. Our results showed that single-environment analysis produced moderate-to-high genomic prediction accuracy for agronomic traits in sesame. The multi-environment analysis further enhanced this accuracy by exploiting marker-by-environment interaction. We concluded that genomic prediction using multi-environmental trial data could improve efforts for breeding cultivars adapted to the semi-arid Mediterranean climate.

Introduction

Sesame (*Sesamum indicum* L.) is an ancient oilseed crop with an annual global production of 6.8 million tons (`https://www.fao.org/faostat/en/#data/QCL`), and there is an increasing demand for its consumption because of its valuable nutritional components. Sesame seeds are rich in high-quality fatty acids, proteins, minerals, and antioxidants, which have health benefits (Wei et al., 2022). The recent availability of sesame genome resources (Berhe et al., 2021; Wang et al., 2022) has provided an opportunity for quantitative genetic modeling of sesame populations. For example, using these resources, quantitative trait loci mapping and genome-wide association analysis in sesame have been conducted for identifying its morphological traits (Mei et al., 2017; Sabag et al., 2021), yield components (Zhou et al., 2018; Sabag et al., 2021), plant architecture (Teboul et al., 2022), response to biotic (Asekova et al., 2021) and abiotic (Li et al., 2018; Dossa et al., 2019) stresses, and seed quality traits (Teboul et al., 2020; Cui et al., 2021) to understand the underlying genetic basis. However, little is known regarding the ability of genomics to predict genetic or breeding values in sesame. Complex traits are influenced by multiple genes, with small effects that are not statistically significant. To address this challenge, genomic predictions that simultaneously accommodate all available genetic markers in regression models to predict genetic or breeding values for capturing marker genetic effects across the whole-genome (Meuwissen et al., 2001) are being used. Genetic or breeding values of lines can be incorporated into selection indices to make a selection decision in breeding (Smith, 1936; Hazel, 1943).

Agronomic traits are influenced by genetic by environment interactions (G $\times$ E) (Gadri et al., 2020). The impact of G $\times$ E ranges from changes in the relative ranking of genotypes to the genomic prediction accuracy, making breeding decisions challenging. With the availability of whole-genome data, the factors of G $\times$ E can be reparametrized as functions of molecular genetic markers via marker-by-environment interactions (M $\times$ E). Recent efforts have included the use of M $\times$ E in whole-genome regression models (Lopez-Cruz et al.,

2015; Crossa et al., 2016). These studies showed that modeling M × E could increase the prediction accuracy compared with that of models without the M × E term.

In this study, we used phenotypic and genomic data from a sesame diversity panel (SCHUJI panel) that was grown over two years (environments) under Mediterranean climatic conditions. This panel was recently used to perform genome-wide association analysis and estimate genomic heritability and genomic correlations for various agronomic traits (Sabag et al., 2021). Our study aimed to evaluate the utility of genomic prediction in predicting sesame traits for both single- and multi-environment analyses.

Materials and Methods

Plant materials, field experiments and genomic data

The complete dataset included phenotypic and genomic data of 182 sesame genotypes from the SCHUJI panel grown over two seasons (2018 and 2020) at the experimental farm of the Hebrew University of Jerusalem (Rehovot, Israel) (Sabag et al., 2021). This panel was characterized by nine agronomic traits: flowering date (FD, in days), height to the first capsule (HTFC, in cm), plant height (PH, in cm), reproductive zone (RZ, in cm), reproductive index (RI, a ratio), number of branches per plant (NBPP), seed-yield per plant (SYPP, g), seed number per plant (SNPP, in gm), and thousand-seed weight (TSW, in gm). The summary statistics for these traits are presented in Table S1. The best linear unbiased estimates of the genotypes were calculated per year by treating the block effect as random (Sabag et al., 2021). Genotyping by sequencing was used to obtain marker information for the 182 genotypes (Elshire et al., 2011). The quality control step included removing tightly linked markers ($r^2 \geq 0.99$), minor allele frequencies less than 0.05, and heterozygosity rates greater than 0.2. The remaining 20,294 single nucleotide polymorphism (SNPs) markers were used for subsequent analyses (Sabag et al., 2021).

Statistical analyses

Single-environment analysis

A single-environment analysis was conducted by fitting two kernel-based methods, genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008) and reproducing kernel Hilbert spaces regression (RKHS) (de los Campos et al., 2010); and two variable selection methods, BayesB (Meuwissen et al., 2001) and BayesC (Kizilkaya et al., 2010).

The kernel-based methods GBLUP and RKHS were fitted as follows.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}$ is the vector of phenotypes; $\mathbf{1}$ is the vector of ones; μ is the overall mean; $\mathbf{Z}$ is the incidence matrix for the random effect; $\mathbf{u} \sim N(0, \mathbf{K}\sigma_u^2)$ is the vector of random genotypes; and $\boldsymbol{\epsilon} \sim$ N $(0,\ \mathbf{I}\sigma_\epsilon^2)$ is the random residual effect. Here, the kernel matrix $\mathbf{K}$ was set to the genomic relationship matrix ($\mathbf{G}$) and the Gaussian kernel matrix ($\mathbf{GK}$) in GBLUP and RKHS, respectively; $\mathbf{I}$ is the identity matrix; σ_u^2 is the genetic variance; and σ_ϵ^2 is the residual variance. The genomic relationship matrix captures additive gene action. In contrast, the Gaussian kernel is equivalent to a space continuous version of the diffusion kernel deployed on graphs (Morota et al., 2013), which can model additive by additive epistatic gene action up to an infinite order (Jiang and Reif, 2015). In GBLUP, $\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{m}$, where $\mathbf{W}$ is a centered and standardized gene content matrix and m is the total number of SNP markers. The Gaussian kernel between a pair of lines i and i' with their marker vectors $\mathbf{w}_i$ and $\mathbf{w}_i'$ is given by

$$\begin{aligned}\mathbf{GK}(\mathbf{w}_i, \mathbf{w}_{i'}) &= \exp(-\theta d_{ii'}^2) \\ &= \prod_{k=1}^{m} \exp(-\theta(w_{ik} - w_{i'k})^2),\end{aligned}$$

where $d_{ii'} = \sqrt{(w_{i1} - w_{i'1})^2 + \cdots + (w_{ik} - w_{i'k})^2 + \cdots + (w_{im} - w_{i'm})^2}$ is the Euclidean distance and θ is the bandwidth parameter. Here, large θ leads to $\mathbf{GK}$ entries closer to 0 (i.e., local kernel), and smaller θ produces entries closer to 1 (i.e., global kernel), controlling the magnitude of genetic similarity between lines. The bandwidth parameter was determined using kernel averaging or multiple kernel learning (de los Campos et al., 2010) by fitting two contrasting kernel matrices with $\theta = 0.2$ and 1.2.

The variable selection methods BayesC and BayesB followed

$$y_i = \mu + \sum_{j=1}^{m} w_{ij}\alpha_j + \epsilon_i, \tag{2}$$

where y_i is the vector of phenotypes for the ith genotype; μ is the overall mean; w_{ij} is the marker covariate at the jth SNP marker coded as 0, 1, or 2; m is the number of SNPs; and α_j is the $j$the marker effect. The prior of α_j for BayesB was:

$$\alpha_j|\pi, \sigma^2_{\alpha_j} = \begin{cases} 0 & \text{with probability of } \pi \\ \sim N(0, \sigma^2_{\alpha_j}) & \text{with probability } (1-\pi) \end{cases}$$

where $\sigma^2_{\alpha_j}$ is the marker genetic variance for the jth SNP and π is a mixture proportion set to 0.99. A Gaussian prior $N(0, \sigma^2_\epsilon)$ was assigned to the vector of residuals, and a flat prior was assigned to μ. The scaled inverse χ^2 distribution was assigned to $\sigma^2_{\alpha_j}$ by setting the degrees of freedom equal to 5 and choosing the scale parameter, assuming that the model explained 50% of the phenotypic variance. In BayesC, $\sigma^2_{\alpha_j}$ was replaced with the common marker genetic variance σ^2_α.

Multi-environment analysis

A multi-environment analysis was conducted using the M × E model (Lopez-Cruz et al., 2015). The core idea of the M × E model is to partition the total marker genetic effects into the main marker genetic effects across all environments and specific marker effects in each environment. As a vector of genetic values consists of a linear combination of marker effects, G × E GBLUP is equivalent to M × E ridge regression BLUP (RR-BLUP). The M × E RR-BLUP model is expressed as $y_{il} = \mu_l + \sum_{k=1}^{m} w_{ilk}(\alpha_{0k} + \alpha_{lk}) + \epsilon_{il}$, where $\alpha_{0\cdot}$ is the main effect of the markers stable for all the environments, $\alpha_{l\cdot}$ is the specific effect of the markers

unique for each environment, and l is the lth environment. In matrix notation,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1}\boldsymbol{\mu}_1 \\ \mathbf{1}\boldsymbol{\mu}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \boldsymbol{\beta}_0 + \begin{bmatrix} \mathbf{W}_1 & 0 \\ 0 & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{bmatrix}$$

where $\begin{bmatrix} \mathbf{1}\boldsymbol{\mu}_1 \\ \mathbf{1}\boldsymbol{\mu}_2 \end{bmatrix}$ is the vector of grand means; $\begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix}$ is the matrix of centered and standardized marker matrix for each environment; $\boldsymbol{\beta}_0 \sim N(0, \mathbf{I}\sigma^2_{\beta_0})$ is the marker effects among environments; the variance component $\sigma^2_{\beta_0}$ is common across the environments and borrows information among them; $\boldsymbol{\beta}_1 \sim N(0, \mathbf{I}\sigma^2_{\beta_1})$ and $\boldsymbol{\beta}_2 \sim N(0, \mathbf{I}\sigma^2_{\beta_2})$ capture the environment specific marker effects with their environment specific variances; and $\boldsymbol{\epsilon}_1 = N(0, \mathbf{I}\sigma^2_{\epsilon_1})$ and $\boldsymbol{\epsilon}_2 = N(0, \mathbf{I}\sigma^2_{\epsilon_2})$ are the heterogeneous residual variances. The extent of variance components associated with $\boldsymbol{\beta}_0$ relative to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ suggests the importance of M $\times$ E. The grand mean was assigned a flat prior. The variance components of markers were drawn from a scaled inverse χ^2 distribution with degrees of freedom $\nu = 5$ and scale parameter s such that the prior means of variance components equal half of the phenotypic variance.

Additionally, the genomic correlation between the same trait in different environments was estimated using a bivariate GBLUP model by extending the single-environment variance-covariance structure to

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{u}} \otimes \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \mathbf{I} \end{pmatrix} \right],$$

where $\mathbf{I}$ is an identity matrix and $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are genetic and residual variance-covariance matrices, respectively. Genomic correlations were derived as $\frac{\sigma^2_{u^*_1 u^*_2}}{\sqrt{\sigma^2_{u^*_1}}\sqrt{\sigma^2_{u^*_2}}}$ where $\sigma^2_{u^*_1 u^*_2}$ is the additive genetic covariance of the trait between the two environments, and $\sigma^2_{u^*_1}$ and $\sigma^2_{u^*_2}$ are additive genetic variances of the trait in 2018 and 2020, respectively. The covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, were assigned an inverse Wishart prior distribution with $\mathbf{W}^{-1}(\mathbf{S}_u, \nu_u)$

and $\mathbf{W}^{-1}(\mathbf{S}_\epsilon, \nu_\epsilon)$, respectively; $\mathbf{S}_u$ and $\mathbf{S}_\epsilon$ are the identity matrices; and ν_u and ν_ϵ are the degrees of freedom. In addition, the phenotypic correlation between the two environments was estimated using the sample phenotypic correlation and the variance components obtained from the M $\times$ E model. The full data set was used to estimate the variance components and genetic correlations.

All the models were implemented in a Bayesian manner. Posterior inferences were based on 50,000 Markov chain Monte Carlo samples, 20,000 burn-in, and a thinning rate of 5 using the BGLR R package following default rules for choices of hyperparameters (Pérez and de Los Campos, 2014; Pérez-Rodríguez and de Los Campos, 2022).

Cross-validation scenarios

For the single-environment analysis, the prediction accuracies of the GBLUP, BayesB, BayesC, and RKHS models were evaluated using the repeated random subsampling cross-validation (CV) (Figure 1). Two-thirds of the lines were used as a training set (TRN) and the remaining one-third were used as a testing set (TST). We measured the predictive Pearson correlation for each repeat, between the observed and predicted values in the TST. The average across 50 replications was used to derive the prediction accuracy of the model.

The predictive ability of the multi-year analysis was assessed using three different CV scenarios that simulated various prediction challenges in plant breeding (Burgueño et al., 2012) (Figure 1). In the first scenario, leave one environment-out CV (CV0), used all the lines in one environment to predict the same lines in a new environment. The second scenario (CV1) predicted the performance of new lines that were not phenotyped in either environment. This scenario evaluated whether newly developed lines that had never been observed in any of the environments could be predicted from their genetic relationships with other lines. In this scenario, the same lines in the same environments were used as TRN, whereas the remaining lines were used for TST. The third CV scenario (CV2) posed the following challenge: some lines were evaluated in only one environment owing to the sparse

field design. In this case, the prediction leveraged both genetic and environmental relationships. The GBLUP model was used to evaluate CV0, and the performance of the M × E RR-BLUP model was benchmarked with that of GBLUP in CV1 and CV2. The repeated random subsampling CV was employed for CV1 and CV2.

Data availability

The phenotypic and genomic information can be found at https://figshare.com/s/94a222afca9423d0b1aa and https://figshare.com/s/a061d548a97237b51a61, respectively.

Results

The sample phenotypic correlations between the environments were all positive, ranging from 0.50 (SNPP) to 0.96 (FD) (Table 1). Similarly, variance component-derived phenotypic correlations were all positive, ranging from 0.37 (SNPP) to 0.80 (FD) (Table 1). Genomic correlation estimates between the environments were all positive, ranging from 0.63 (SNPP) to 0.97 (FD) (Table 1).

Single-environment genomic prediction

Single-environment prediction accuracies of the nine agronomic traits were evaluated using the four whole-genome regression models (Figure 2 and Table S2). Overall, no notable difference was observed between the environments and the models. The highest mean prediction accuracy was obtained for HTFC (0.77 and 0.78 in 2018 and 2020, respectively, averaged across the models), whereas the lowest was for SNPP in 2018 (0.49) and SYPP in 2020 (0.39). FD, PH, RI, and NBPP showed relatively high prediction accuracies. In particular, the prediction accuracies ranged from 0.74 in 2018 to 0.70 in 2020 for FD, 0.68 in 2018 to 0.67 in 2020 for PH, 0.71 in 2018 to 0.74 in 2020 for RI, and 0.69 in 2018 to 0.62 in 2020 for NBPP. The prediction accuracy of RZ was slightly lower than that of these traits, with 0.56 in 2018 and 0.53 in 2020. The three yield-related traits SYPP, SNPP and TSW showed moderate prediction accuracies of 0.57 and 0.39, 0.49 and 0.40, and 0.55 and 0.50 for 2018 and 2020, respectively. The prediction accuracies for 2018 were higher than those for 2020.

Multi-environment genetic parameter estimation

Variance component estimates were obtained from the M $\times$ E RR-BLUP model using the full data set and expressed in terms of proportions (Figure 3). In the two yield-related traits, SYPP and SNPP, the M $\times$ E components were largest whereas the additive genetic

components were the lowest. However, the extent of M × E was lower for FD, HTFC, RI, and TSW. Similarly, the estimates of genomic heritability were low for SYPP and SNPP, and high for FD, HTPC, RI, and TSW (Table 1). Estimates of genomic correlations between the two environments were all moderate to high, ranging from 0.63 (SNPP) to 0.97 (FD) (Table 1).

Multi-environment genomic prediction

One of the main challenges for the genomic prediction of multi-environmental data was predicting the performance of new or observed lines in new or known environments. We used multi-environment data to evaluate the genomic prediction accuracies of nine important agronomic traits in sesame by accounting for M × E. Our main objective was to investigate whether obtaining information from another environment could improve predictions compared to a single-environment analysis. As we did not observe a difference among GBLUP, BayesB, BayesC, and RKHS in the single-environment analysis, multi-environment analysis was conducted using the GBLUP or RR-BLUP type of models.

CV0 scenario: In the CV0 scenario, all lines in one environment were used to predict the same lines in a new environment by applying the GBLUP model (Figure 1B). Overall, we obtained an improvement in the prediction accuracies of all traits compared to the single-environment model (Figure 4). The prediction accuracies were highest for FD and HTFC, with 0.93 and 0.92, respectively. For other agronomic traits, the prediction accuracies ranged between 0.78 (NBPP) and 0.9 (RI). For yield components, prediction accuracies were 0.63, 0.55, and 0.74 for SYPP, SNPP, and TSW, respectively.

CV1 scenario: The CV1 scenario mimicked the situation in which we aimed to predict the performance of new lines (Figure 1C). We did not observe a major difference between the single-environment and M × E models (Figure 5 and Supplemental Table S3). The prediction accuracies from multi-environment analysis were almost equal to or lower than those from the single-environment analysis for some traits.

CV2 scenario: In this scenario, we evaluated the multi-environment analysis when some of the lines were not evaluated in all environments (Figure 1D). Large improvements were observed for all traits (Figure 5). The predictive accuracies of CV2 were greater than those of CV1 and the single environment GBLUP. For 2018 and 2020, improvements ranged from 17% (HTFC) to 48% (TSW) and from 15% (HTFC) to 58% (TSW), respectively. The differences in improvements were statistically significant (Table S3). Although the single-environment prediction accuracies of the yield-related traits, SYPP and SNPP, were low, using the M $\times$ E model, the gains achieved were 20% and 45% for 2018 and 20% and 28% for 2020, respectively, compared to those obtained from the single-environment analysis.

Discussion

The future of food systems and security relies heavily on accelerating plant breeding (Lenaerts et al., 2019). Developing new varieties with high nutritional value and integration of Orphan crops such as sesame provide new opportunities to expend the human diet quality and sustainability (Dawson et al., 2019). Among the modern methods for plant breeding, genomic selection has proven effective in terms of genetic gain (Voss-Fels et al., 2019). In this study, we evaluated the genomic prediction accuracies of nine agronomic traits in sesame using a diversity panel. This was the first critical step taken toward establishing a genomic selection program for sesame.

Performance of single-environment genomic prediction

Overall, we observed moderate-to-high prediction accuracies for all traits in the single-environment analysis (Figure 2). We did not find any significant differences between GBLUP, BayesB, BayesC, and RKHS. Variable selection methods, such as BayesB and BayesC, are expected to perform better than GBLUP in the presence of large quantitative trait locus effects (Daetwyler et al., 2010). Comparable prediction performance between GBLUP and variable selection methods supported a previous genome-wide association study reporting that only a few significant loci influenced the studied traits using the same sesame panel (Sabag et al., 2021). This suggests that agronomic traits in sesame are mostly controlled by many small-effect quantitative trait loci rather than by major quantitative trait loci. In addition, we found an association between the genomic heritability estimates and prediction accuracy. The higher the genomic heritability estimate, the higher the accuracy of genomic prediction. For example, FD and HTFC showed high genomic heritability estimates (0.72 and 0.68, respectively) and high prediction accuracies (0.72 and 0.78 on average, respectively, for both environments). Similarly, the yield components SYPP and SNPP had the lowest

prediction accuracies in the two environments, as well as the lowest genomic heritability estimates. Numerous factors affect genomic prediction accuracy, such as genetic architecture, the quantitative genetic model used, trait heritability, marker density, size of the reference population, and the genetic relationship between TRN and TST (Daetwyler et al., 2010). For example, given the small sample size of the sesame diversity panel (Sabag et al., 2021), increasing the number of lines could improve the predictive performance of lowly heritable traits, such as yield components (e.g., SYPP and SNPP).

Multi-environment analysis to enhance genomic prediction

Understanding genotype-by-environment interactions are among the main challenges for plant breeding (Cooper and DeLacy, 1994; Mathews et al., 2008). The M × E model decomposes the marker effect into the marker main effect, which borrows information from the other environment, and the marker-specific effect for each environment (Lopez-Cruz et al., 2015). No notable improvement from the M × E model was observed for CV1 when predicting the performance of new lines that were not observed in any environment. This agreed with previous reports of no strong evidence of gain in prediction for the CV1 scenario using the M × E model compared to single-environment analysis (Burgueño et al., 2012; Lopez-Cruz et al., 2015; Crossa et al., 2016). In this scenario, no information was borrowed from the other environment. In such a case, integrating environmental covariates into the prediction model may be an alternative strategy for improving the prediction accuracy (Jarquín et al., 2014).

Many lines are often evaluated simultaneously for multiple environments in plant breeding programs (Lorenz, 2013). This leads to unbalanced field experimental designs (Lado et al., 2016), in which not all lines are present in all environments. We simulated this scenario using CV2 to investigate whether capturing environmental information improved the prediction

accuracies of agronomic traits in sesame. In general, considerable improvement in prediction accuracies were observed with the M × E model compared to those of GBLUP for all traits in all environments. Our results concurred with those of previous studies (Lopez-Cruz et al., 2015; Crossa et al., 2016; Cuevas et al., 2016; Bandeira e Sousa et al., 2017; Cuevas et al., 2018), suggesting that the M × E model borrowed environmental information across environments and improved prediction accuracies (Lopez-Cruz et al., 2015). In particular, the M × E model performed well when the sample phenotypic correlations between environments were positive (Lopez-Cruz et al., 2015). This is because the covariance between any two environments is linearly related to the proportion of the genetic variance, explained by the marker main effect in the M × E model, causing the phenotypic correlation between the two environments to be positive or zero in our data. The pairs of phenotypic correlations between the environments were positive for all the agronomic traits. The mean (standard deviation) of the sample phenotypic correlation between the environments was 0.79 (0.16) (Table 1). This led to a correlation between the sample- and the ratio of variance component-based phenotypic correlations of 0.95. The positive sample phenotypic correlation between the two environments might be a critical factor in explaining why the M × E model outperformed the single-environment GBLUP model in CV2. In addition, the largest gain in prediction in CV0 compared to that in the single-environment analysis was achieved for traits with a large extent of M × E components (SNPP and SYPP) (Table 1 and Figure 4). This finding indicated that when G × E is present, the M × E model can improve prediction accuracy. Although we employed the M × E model, which only captured additive genetic effects, the extension of G × E GBLUP to RKHS has been reported to outperform G × E GBLUP in maize and wheat grain yield, especially when many environments were analyzed (Cuevas et al., 2016).

The future of genomic prediction in a sesame breeding

Crop rotation is critical for sustainable agricultural production systems (Li et al., 2019), and the introduction of new crops, such as sesame, can be used for this purpose. Although sesame is primarily cultivated in developing countries with relatively low yields (Dossa et al., 2017), its demand for consumption is increasing. Accelerated breeding efforts are necessary to meet this growing demand. In this study, we performed genomic prediction for nine important agronomic traits in sesame using single- and multi-environment analyses for the first time. As genomic prediction is an essential first step toward the implementation of genomic selection in breeding programs, we examined the potential of using genomic prediction to enhance genetic gain in sesame while accounting for M $\times$ E. Additional improvements in yield components may be achieved using a multi-trait model along with secondary traits evaluated in this study or applying high-throughput phenotyping during the growing season (Morota et al., 2022).

Conclusions

Currently, genetic research on sesame is limited to quantitative trait locus mapping (Teboul et al., 2020) or genome-wide association studies (Berhe et al., 2021; Sabag et al., 2021). In this study, we evaluated the usefulness of whole-genome prediction models in predicting important agronomic traits in sesame. Overall, we obtained moderate-to-high genomic prediction accuracies. Prediction performance was further enhanced by accounting for M $\times$ E. Given the reduced cost of genotyping and the availability of high-quality genomic resources for sesame, we conclude that genomic prediction has the potential to facilitate sesame breeding by transforming the prediction gain into selection decisions in Mediterranean climatic conditions.

Author contribution statement

IS and ZP performed the field experiments. IS analyzed the data. IS drafted the manuscript. YB and GM supported IS on the data analysis. YB, ZP, and GM edited the manuscript. ZP and GM supervised the study.

Acknowledgments

This research was supported by a Research Grant from BARD, the United States - Israel Binational Agricultural Research and Development Fund (No. IS-5400-21), the Hebrew University of Jerusalem, and Virginia Polytechnic Institute and State University. I.S. is indebted to the Samuel and Lottie Rudin scholarship foundation.

References

Asekova, S., Oh, E., Kulkarni, K. P., Siddique, M. I., Lee, M. H., Kim, J. I., Lee, J.-D., Kim, M., Oh, K.-W., Ha, T.-J., et al. (2021). An integrated approach of QTL mapping and genome-wide association analysis identifies candidate genes for phytophthora blight resistance in sesame (sesamum indicum l.). *Frontiers in Plant Science*, 12.

Bandeira e Sousa, M., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., Burgueño, J., and Crossa, J. (2017). Genomic-enabled prediction in maize using kernel models with genotype× environment interaction. *G3: Genes, Genomes, Genetics*, 7(6):1995–2014.

Berhe, M., Dossa, K., You, J., Mboup, P. A., Diallo, I. N., Diouf, D., Zhang, X., and Wang, L. (2021). Genome-wide association study and its applications in the non-model crop *Sesamum indicum*. *BMC Plant Biology*, 21(1):1–19.

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, 52(2):707–719.

Cooper, M. and DeLacy, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88(5):561–572.

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., and Pérez-Rodríguez, P. (2016). Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Science*, 56(5):2193–2209.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., Campos, G. d. l., Montesinos-López, O., and Burgueño, J. (2016). Genomic prediction

of genotype× environment interaction kernel regression models. *The Plant Genome*, 9(3):plantgenome2016–03.

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M., and Crossa, J. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. *G3: Genes, Genomes, Genetics*, 8(4):1347–1365.

Cui, C., Liu, Y., Liu, Y., Cui, X., Sun, Z., Du, Z., Wu, K., Jiang, X., Mei, H., and Zheng, Y. (2021). Genome-wide association study of seed coat color in sesame (sesamum indicum l.). *Plos One*, 16(5):e0251526.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3):1021–1031.

Dawson, I. K., Powell, W., Hendre, P., Bančič, J., Hickey, J. M., Kindt, R., Hoad, S., Hale, I., and Jamnadass, R. (2019). The role of genetics in mainstreaming the production of new and orphan crops to diversify food systems and support human nutrition. *New Phytologist*, 224(1):37–54.

de los Campos, G., Gianola, D., Rosa, G. J., Weigel, K. A., and Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. *Genetics Research*, 92(4):295–308.

Dossa, K., Diouf, D., Wang, L., Wei, X., Zhang, Y., Niang, M., Fonceka, D., Yu, J., Mmadi, M. A., Yehouessi, L. W., et al. (2017). The emerging oilseed crop sesamum indicum enters the "omics" era. *Frontiers in Plant Science*, 8:1154.

Dossa, K., Li, D., Zhou, R., Yu, J., Wang, L., Zhang, Y., You, J., Liu, A., Mmadi, M. A., Fonceka, D., et al. (2019). The genetic basis of drought tolerance in the high oil crop sesamum indicum. *Plant Biotechnology Journal*, 17(9):1788–1803.

372 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and
373 Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (gbs) approach for high
374 diversity species. *PloS one*, 6(5):e19379.

375 Gadri, Y., Williams, L. E., and Peleg, Z. (2020). Tradeoffs between yield components promote
376 crop stability in sesame. *Plant Science*, 295:110105.

377 Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28(6):476–
378 490.

379 Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux,
380 F., Guerreiro, L., Pérez, P., Calus, M., et al. (2014). A reaction norm model for genomic
381 selection using high-dimensional genomic and environmental data. *Theoretical and Applied*
382 *Genetics*, 127(3):595–607.

383 Jiang, Y. and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics*,
384 201(2):759–768.

385 Kizilkaya, K., Fernando, R., and Garrick, D. (2010). Genomic prediction of simulated multi-
386 breed and purebred performance using observed fifty thousand single nucleotide polymor-
387 phism genotypes. *Journal of animal science*, 88(2):544–551.

388 Lado, B., Barrios, P. G., Quincke, M., Silva, P., and Gutiérrez, L. (2016). Modeling
389 genotype× environment interaction for genomic selection with unbalanced data from a
390 wheat breeding program. *Crop Science*, 56(5):2165–2179.

391 Lenaerts, B., Collard, B. C., and Demont, M. (2019). Improving global food security through
392 accelerated plant breeding. *Plant Science*, 287:110207.

393 Li, D., Dossa, K., Zhang, Y., Wei, X., Wang, L., Zhang, Y., Liu, A., Zhou, R., and Zhang,
394 X. (2018). GWAS uncovers differential genetic bases for drought and salt tolerances in
395 sesame at the germination stage. *Genes*, 9(2):87.

Li, J., Huang, L., Zhang, J., Coulter, J. A., Li, L., and Gan, Y. (2019). Diversifying crop rotation improves system robustness. *Agronomy for Sustainable Development*, 39(4):1–13.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., Singh, R. P., Autrique, E., and de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker× environment interaction genomic selection model. *G3: Genes, Genomes, Genetics*, 5(4):569–582.

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

Mathews, K. L., Malosetti, M., Chapman, S., McIntyre, L., Reynolds, M., Shorter, R., and Van Eeuwijk, F. (2008). Multi-environment QTL mixed models for drought stress adaptation in wheat. *Theoretical and Applied Genetics*, 117(7):1077–1091.

Mei, H., Liu, Y., Du, Z., Wu, K., Cui, C., Jiang, X., Zhang, H., and Zheng, Y. (2017). High-density genetic map construction and gene mapping of basal branching habit and flowers per leaf axil in sesame. *Frontiers in Plant Science*, 8:636.

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

Morota, G., Jarquin, D., Campbell, M. T., and Iwata, H. (2022). Statistical methods for the quantitative genetic analysis of high-throughput phenotyping data. In *High-Throughput Plant Phenotyping*, pages 269–296. Springer.

Morota, G., Koyama, M., M Rosa, G. J., Weigel, K. A., and Gianola, D. (2013). Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genetics Selection Evolution*, 45(1):1–15.

Pérez, P. and de Los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495.

Pérez-Rodríguez, P. and de Los Campos, G. (2022). Multitrait bayesian shrinkage and variable selection models with the bglr-r package. *Genetics*, 222(1):iyac112.

Sabag, I., Morota, G., and Peleg, Z. (2021). Genome-wide association analysis uncovers the genetic architecture of tradeoff between flowering date and yield components in sesame. *BMC Plant Biology*, 21(1):1–14.

Smith, H. F. (1936). A discriminant function for plant selection. *Annals of eugenics*, 7(3):240–250.

Teboul, N., Gadri, Y., Berkovich, Z., Reifen, R., and Peleg, Z. (2020). Genetic architecture underpinning yield components and seed mineral–nutrients in sesame. *Genes*, 11(10):1221.

Teboul, N., Magder, A., Zilberberg, M., and Peleg, Z. (2022). Elucidating the pleiotropic effects of sesame kanadi1 locus on leaf and capsule development. *The Plant Journal*, 110(1):88–102.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, 132(3):669–686.

Wang, M., Huang, J., Liu, S., Liu, X., Li, R., Luo, J., and Fu, Z. (2022). Improved assembly and annotation of the sesame genome. *DNA Research*.

Wei, P., Zhao, F., Wang, Z., Wang, Q., Chai, X., Hou, G., and Meng, Q. (2022). Sesame (sesamum indicum l.): A comprehensive review of nutritional value, phytochemical composition, health benefits, development of food, and industrial applications. *Nutrients*, 14(19):4079.

443 Zhou, R., Dossa, K., Li, D., Yu, J., You, J., Wei, X., and Zhang, X. (2018). Genome-
444 wide association studies of 39 seed yield-related traits in sesame (sesamum indicum l.).
445 *International Journal of Molecular Sciences*, 19(9):2794.

Tables

Trait	h^2	r_g	r_y	r'_y
FD	0.72	0.97	0.96	0.80
HTFC	0.68	0.94	0.95	0.77
PH	0.57	0.82	0.83	0.66
RZ	0.62	0.87	0.82	0.71
RI	0.68	0.92	0.93	0.75
NBPP	0.55	0.83	0.78	0.65
SYPP	0.38	0.76	0.58	0.47
SNPP	0.29	0.63	0.50	0.37
TSW	0.70	0.87	0.80	0.77

Table 1: Genomic heritability estimates of the nine agronomic sesame traits (h^2), genetic correlations (r_g), sample phenotypic correlations (r_y), and variance-components derived phenotypic correlations (r'_y) between the two environment using the marker-by-environment interaction model. Flowering date (FD), height to the first capsule (HTFC), plant height (PH), reproductive zone (RZ), reproductive index (RI), number of branches per plant (NBPP), seed-yield per plant (SYPP), seeds number per plant (SNPP), and thousand-seed weight (TSW).

Figures

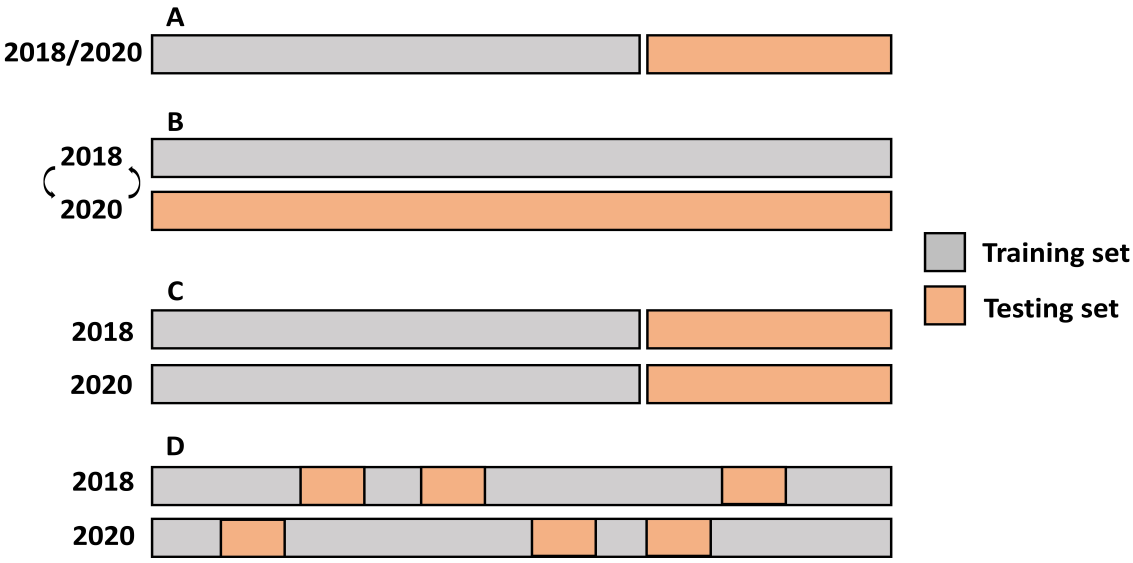

Figure 1: Single- and multi-environment genomic prediction cross-validation scenarios. A: Single-environment analysis, B: All the lines in one environment were used to predict the same lines in a new environment (CV0), C: Performance of new lines that are not phenotyped in any environment was predicted through the genetic relationship with other lines (CV1), and D: Predict lines that were evaluated in only one environment through the genetic and environmental relationships (CV2).

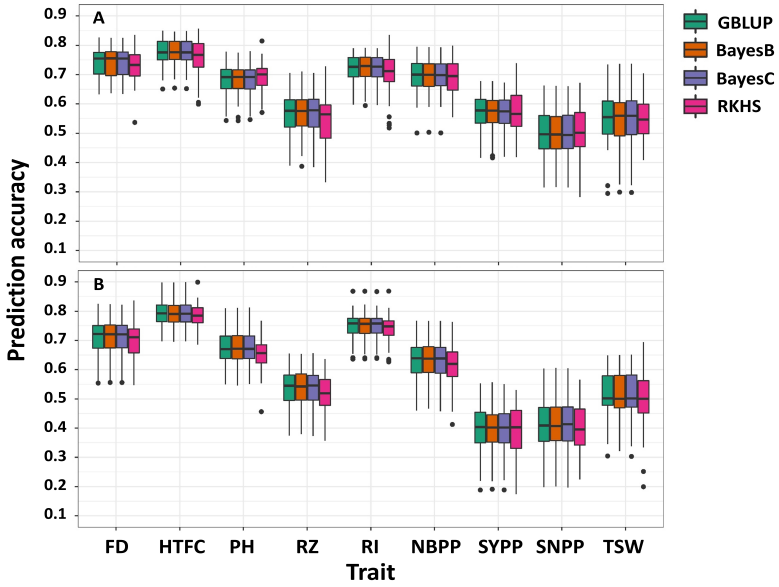

Figure 2: Single-environment prediction accuracies of the nine agronomic sesame traits in 2018 (**A**) and 2020 (**B**) growing seasons using genomic best linear unbiased prediction (GBLUP), BayesB, BayesC, and reproducing kernel Hilbert spaces regression (RKHS). Flowering date (FD), height to the first capsule (HTFC), plant height (PH), reproductive zone (RZ), reproductive index (RI), number of branches per plant (NBPP), seed-yield per plant (SYPP), seeds number per plant (SNPP), and thousand-seed weight (TSW).

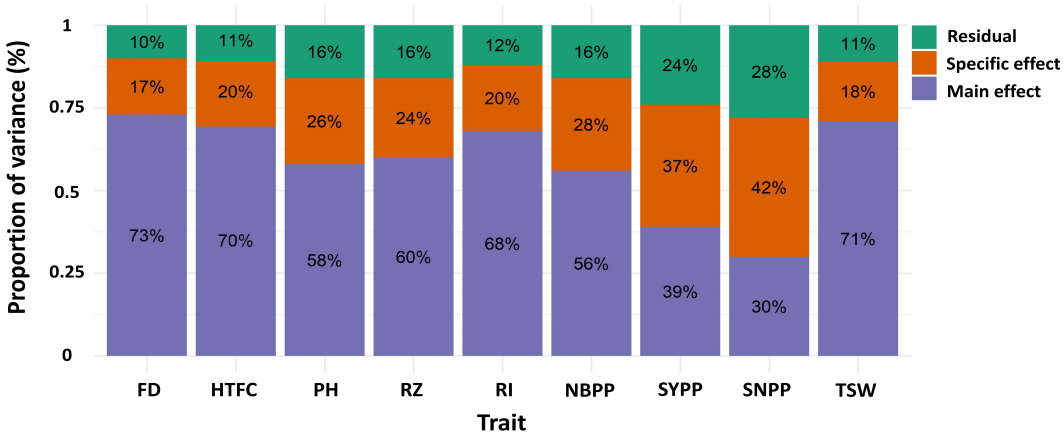

Figure 3: Proportion of the main genetic variance, environment-specific variance, and residual variance components for each trait obtained from the marker-by-environment interaction model. Flowering date (FD), height to the first capsule (HTFC), plant height (PH), reproductive zone (RZ), reproductive index (RI), number of branches per plant (NBPP), seed-yield per plant (SYPP), seeds number per plant (SNPP), and thousand-seed weight (TSW).

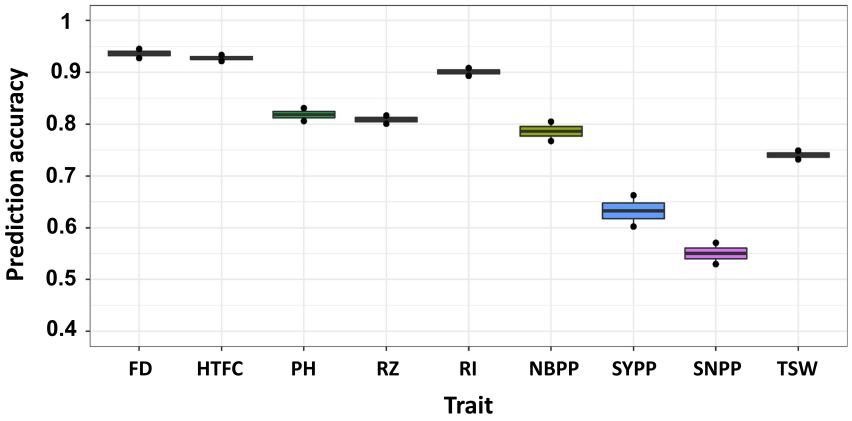

Figure 4: Multi-environment genomic prediction accuracies of the nine agronomic sesame traits using the best linear unbiased prediction model when all the lines in one environment were used to predict the same lines in a new environment (CV0). Flowering date (FD), height to the first capsule (HTFC), plant height (PH), reproductive zone (RZ), reproductive index (RI), number of branches per plant (NBPP), seed-yield per plant (SYPP), seeds number per plant (SNPP), and thousand-seed weight (TSW).

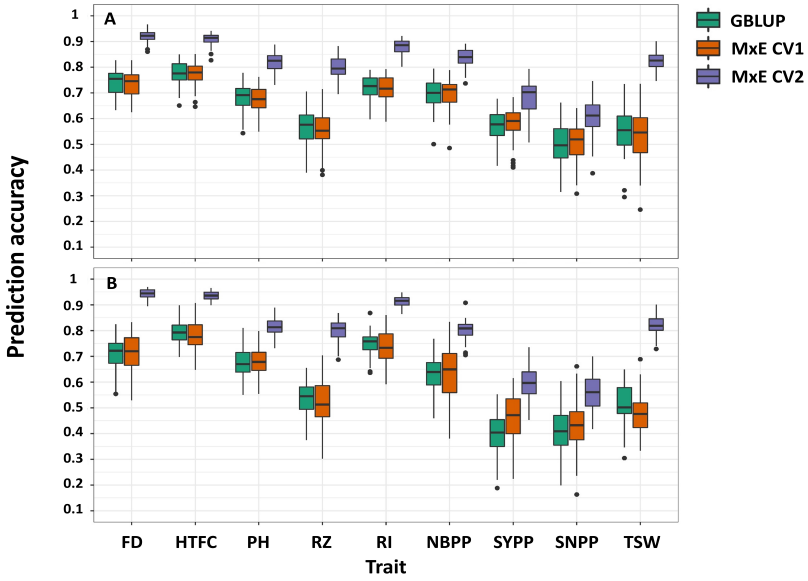

Figure 5: Comparison of prediction accuracies in single- and multi-environment models for predicting new lines that are not phenotyped in any environment (CV1) and predicting lines that were evaluated in only one environment (CV2) in 2018 (A) and 2020 (B) growing seasons. Flowering date (FD), height to the first capsule (HTFC), plant height (PH), reproductive zone (RZ), reproductive index (RI), number of branches per plant (NBPP), seed-yield per plant (SYPP), seeds number per plant (SNPP), and thousand-seed weight (TSW).