the geneticssociety

www.nature.com/hdy

ARTICLE OPEN

Multi-generation genomic prediction of maize yield using parametric and non-parametric sparse selection indices

Marco Lopez-Cruz[1,2]✉, Yoseph Beyene[3], Manje Gowda[3], Jose Crossa[4,5], Paulino Pérez-Rodríguez[5] and Gustavo de los Campos[2,6,7]

Genomic prediction models are often calibrated using multi-generation data. Over time, as data accumulates, training data sets become increasingly heterogeneous. Differences in allele frequency and linkage disequilibrium patterns between the training and prediction genotypes may limit prediction accuracy. This leads to the question of whether all available data or a subset of it should be used to calibrate genomic prediction models. Previous research on training set optimization has focused on identifying a subset of the available data that is optimal for a given prediction set. However, this approach does not contemplate the possibility that different training sets may be optimal for different prediction genotypes. To address this problem, we recently introduced a sparse selection index (SSI) that identifies an optimal training set for each individual in a prediction set. Using additive genomic relationships, the SSI can provide increased accuracy relative to genomic-BLUP (GBLUP). Non-parametric genomic models using Gaussian kernels (KBLUP) have, in some cases, yielded higher prediction accuracies than standard additive models. Therefore, here we studied whether combining SSIs and kernel methods could further improve prediction accuracy when training genomic models using multi-generation data. Using four years of doubled haploid maize data from the International Maize and Wheat Improvement Center (CIMMYT), we found that when predicting grain yield the KBLUP outperformed the GBLUP, and that using SSI with additive relationships (GSSI) lead to 5–17% increases in accuracy, relative to the GBLUP. However, differences in prediction accuracy between the KBLUP and the kernel-based SSI were smaller and not always significant.

Heredity (2021) 127:423–432; https://doi.org/10.1038/s41437-021-00474-1

INTRODUCTION

Almost two decades have passed since Genomic Selection (GS) was first proposed by Meuwissen et al. (2001). This groundbreaking idea was quickly adopted for breeding dairy cattle (Hayes et al. 2009), beef cattle (Garrick 2011), broilers (Wolc et al. 2016), maize (Bernardo and Yu 2007), wheat (Poland et al. 2012), and many other animal and crop species (Xu et al. 2020). Over time, investments by public and private organizations led to the development of large genomic data sets comprising DNA sequences and phenotypes. These large sample sizes of modern genomic data sets have increased our ability to accurately train high-dimensional genomic prediction models and methods (Howard et al. 2019).

However, a larger sample size often comes with increased genetic heterogeneity, including many generations of data and often complex admixture patterns. Moreover, there have been some indications that in genomic prediction, *bigger may not always be better*. For example, Wolc et al. (2016) reported that the accuracy of genomic predictions in a broiler breeding program was higher when using data from the last three generations, relative to prediction equations trained using data from the last five generations. Likewise, Riedelsheimer et al. (2013) and Jacobson et al. (2014) reported that the prediction accuracy was higher when models were trained using data from biparental families that shared at least one parent, relative to training using data from all the available biparental families.

Early work by Habier et al. (2010) showed that family relationships have an important impact on prediction accuracy and many studies have proven that distantly related individuals make a small (sometimes negligible) contribution to the prediction accuracy. However, as noted above, some evidence suggests that using training sets formed by individuals distantly related to the genotypes of the prediction set may actually have a negative impact on the prediction accuracy (e.g., Lorenz and Smith 2015). This may happen if, for example, heterogeneity in allele frequency and in linkage disequilibrium (LD) patterns between the training and prediction sets lead to SNP-effect heterogeneity.

Issues related to data- and effect-heterogeneity have spawned multiple research efforts. One line of research models heterogeneity of effects by explicitly using SNP-by-group interaction models or multivariate models ('multi-breed genomic prediction'), in which effects are assumed to be correlated among groups (e.g., Olson et al. 2012; Lehermeier et al. 2015; Rio et al. 2020).

[1]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA. [2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA. [3]Global Maize Program, International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya. [4]Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Mexico, Mexico. [5]Colegio de Postgraduados, Montecillos, Edo. de México, Mexico. [6]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA. [7]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI, USA. Associate editor: Chenwu Xu. ✉email: lopezcru@msu.edu

Received: 19 April 2021 Revised: 10 September 2021 Accepted: 11 September 2021
Published online: 25 September 2021

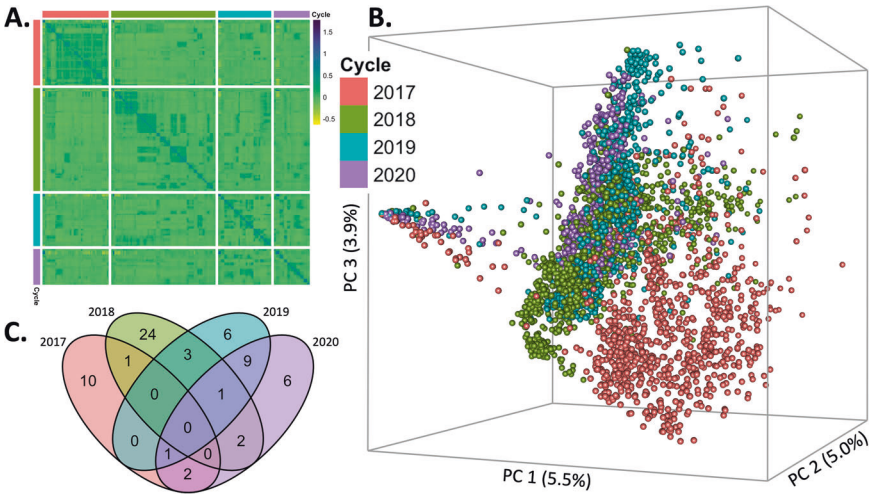

Fig. 1 Genetic structure of the multi-generation maize data. A Heatmap of the genomic relationships matrix. **B** First three principal components of the additive genomic relationships matrix, **G**. Dots represent individuals that are separated by colors for each cycle (2017, 2018, 2019 or 2020). **C** Venn diagram representing the number of common parents used to generate the DH lines at each year.

This approach has shown promise, yet it is only adequate when genotypes can be clustered into clearly disjointed groups.

Another line of research attempts to increase prediction accuracy with the optimal design of training data sets. The methods proposed and used to identify an optimal training set span from simple threshold-based methods (e.g., Clark et al. 2012; Lorenz and Smith 2015) to more sophisticated algorithms that seek to minimize prediction error variance or maximize the expected reliability (e.g., Rincent et al. 2012; Akdemir and Isidro-Sanchez 2019; Roth et al. 2020). A main assumption of these training set optimization methods is that a single training set is optimal for all individuals in the prediction set. However, this may not be the case if some individuals in the training set can improve prediction accuracy for some of the selection candidates and reduce it for others.

To address the limitations of existing methods, Lopez-Cruz and de los Campos (2021) proposed a prediction method (sparse selection index, SSI) that identifies a customized training set for each individual in the prediction set. The SSI integrates into the selection index methodology (Smith 1936; Hazel 1943) a sparsity-inducing penalty that leads to sparse selection indices and, applied to two wheat data sets, outperformed the genomic-BLUP (GBLUP; VanRaden 2008) by 5–10%.

Reproducing Kernel Hilbert Spaces (RKHS) regression has shown good predictive performance in many genomic applications (e.g., de los Campos et al. 2010; González-Camacho et al. 2012). The GBLUP is a special case of RKHS regression in which a linear kernel (additive genomic relationships) is used to describe the genetic similarity between genotypes (de los Campos et al. 2009). However, several studies (e.g., Crossa et al. 2010; de los Campos et al. 2010; Morota and Gianola 2014; Bandeira e Sousa et al. 2017; Cuevas et al. 2016, 2017, 2018) have suggested that using non-linear kernels (e.g., Gaussian kernels) may lead to higher genomic prediction accuracy. In a Gaussian kernel, the covariance between genetic values is higher for closely related individuals and drops as genotypes become increasingly distant. The rate at which the prior covariance between genetic values drops is controlled by a bandwidth parameter. Large bandwidth parameter values (that lead to highly local covariances) can be used to derive predictions that are largely dependent on closely related individuals. Thus, there is a clear link between the RKHS with Gaussian kernels and the SSI methodology of Lopez-Cruz and de los Campos (2021). However, the Gaussian kernel does not yield strictly sparse prediction equations.

Therefore, we study whether the SSI can also improve the prediction performance of RKHS regressions with non-linear kernels. The objective of this study is to evaluate the performance of the SSI using additive (GSSI) and non-additive (KSSI) kernels using four-generations (years) of a DH (doubled haploid) maize data set from the ongoing maize breeding program at the International Maize and Wheat Improvement Center (CIMMYT). For several scenarios of training set composition, we compare the prediction accuracy of the BLUP and the SSI with additive and non-additive kernels.

MATERIALS AND METHODS

Genotypes and phenotypic data

The genotypes used in the study consist of 3722 DH lines derived from 54 biparental families. The DH lines were developed at CIMMYT's Maize DH facility at the Agricultural & Livestock Research Organization (KALRO) in Kiboko, Kenya. The biparental families were obtained by crossing elite inbred lines with drought-tolerant lines that were tested for the last four years. Some parents used to generate the DH lines in one year were also used (based on a 100% pedigree similarity) to create the DH lines in other years (see Fig. 1C). There was one parent in common between 2017 and 2018; two common parents between 2017 and 2020; one common parent between 2017, 2019, and 2020; three common parents between 2018 and 2019; two common parents between 2018 and 2020; one common parent between 2018, 2019, and 2020; and finally, nine common parents between 2018 and 2019. The 3722 DH lines were selected from a larger population (based on the results of evaluating germination, good stand, plant type, low ear placement, and well-filled ears) for stage I multi-location yield trials conducted from 2017 to 2020.

Each year, the selected DH lines were crossed (as male) with a single-cross tester (as female) from the complementary heterotic group to generate tree-way hybrids that were evaluated under well-watered (denoted as *optimal*) and *drought* conditions. The number of hybrids (trials) planted in 2017, 2018, 2019, and 2020 were 923 (14), 1423 (34), 722 (17), and 654 (13), respectively; trials were connected by three to six commercial checks, planted in an alpha-lattice design with two replications and evaluated in two well-watered locations and one managed drought stress location during the 2017, 2018, 2019, and 2020 growing seasons. The optimal experiments were conducted during the rainy season, applying supplemental irrigation as needed. The drought experiments were conducted during the dry (rain-free) season, and irrigation was suspended two weeks before flowering and until harvest. Entries were planted in two-row plots, 5 m long, with a spacing of 0.75 m between rows and 0.25 m between hills. Two seeds were initially planted per hill and afterward, three weeks after emergence, one plant was kept per hill to obtain a final plant density of 53333 plants ha^{-1}. Fertilizers were applied at the rate of 60 kg N and 60 kg P_2O_5 ha^{-1}, as recommended for the area. Nitrogen was applied twice: once when planting, and again 6 weeks after emergence. Fields were kept free of weeds by hand weeding. Grain yield (GY, tons ha^{-1}), anthesis date (AD, days) and plant height (PH, cm) traits were recorded. Plots were manually harvested and GY was corrected to a

moisture of 12.5%. AD was measured from planting to the moment in which 50% of the plants shed pollen, and PH was measured between the soil surface and the flag leaf collar on five representative plants in each plot.

Leaf samples were taken from each of the 3722 DH lines and sent to Intertek, Sweden, for DNA extraction. The DNA sample plates were forwarded to the Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA, for genotyping with repetitive sequences (rAmpSeq) as described by Buckler et al. (2016). A distortion segregation analysis was performed to a total of 5465 dominant markers coded as 0 (absence) and 1 (presence) to detect if the segregation patterns deviate from the expected Mendelian ratio of 3:1. A Benjamini–Hochberg False Discovery Rate correction was applied to the *P* values to account for multiple testing; a total of 61 markers were discarded at a 5% FDR. The remaining markers were filtered by minor allele frequency (MAF < 0.05), leading 4612 filtered markers that were used for analyses.

Data from the 2017 and 2018 cycles have been previously described and analyzed by Beyene et al. (2019) and Atanda et al. (2021).

Phenotypes' pre-processing

The adjusted means of GY, AD and PH were obtained using mixed-effects models fitted separately for each trait-environmental-condition-year combination. The Best Linear Unbiased Estimates (BLUE) of genotypes for the optimal experiments were estimated within year across the two locations using the META-R software (Alvarado et al. 2020) following the linear mixed model:

$$Y_{ijkl} = \mu + G_i + L_j + R_{k(j)} + B_{l(kj)} + (G \times L)_{ij} + e_{ijkl}$$

where Y_{ijkl} is the phenotypic record of genotype i in location j in replicate k within block l, μ is the overall mean, L_j is the fixed effect of location j, $R_{k(j)}$ is the fixed effect of the replicate k within location j, $B_{l(kj)}$ is the random effect of the incomplete block l within replicate k and location j assumed to be independently and identically (iid) normal distributed with a mean of zero and a variance of σ^2_b, G_i is the fixed effect of genotype i, $(G \times L)_{ij}$ is the fixed effect of the genotype × location interaction, and e_{ijkl} is the random error assumed to be iid normal with mean zero and variance σ^2_ε. After fitting the model just described, adjusted phenotypes (y_i) were obtained by subtracting the estimated effects of location, replicate, incomplete block, genotype × location interaction and error from each phenotypic record. Likewise, within each year, the BLUE for each trait for the single-location drought experiment was obtained through the linear model

$$Y_{ikl} = \mu + G_i + R_k + B_{l(k)} + e_{ikl}$$

where R_k is the fixed effect of the replicate k, $B_{l(k)}$ is the random effect of the incomplete block l within replicate k assumed to be *iid* normal with a mean of zero and a variance σ^2_b, and the remaining factors are as before. The adjusted phenotypes were obtained by subtracting the estimated effects of replicate, incomplete block, and error from the phenotypic records.

A total of $n = 3527$ lines containing marker information and phenotypic information were kept for GS models. The final number of lines in 2017, 2018, 2019, and 2020 were $n_1 = 901$, $n_2 = 1418$, $n_3 = 722$, and $n_4 = 486$, respectively.

Genomic prediction methods

We considered four prediction models: genomic-BLUP (GBLUP) using additive genomic relationships (VanRaden 2008); Reproducing Kernel Hilbert Spaces (RKHS) regression (Gianola et al. 2006; de los Campos et al. 2010), which is equivalent to a GBLUP with a non-linear kernel (de los Campos et al. 2009); and sparse selection indices (SSI) obtained by imposing an L1-penalty on a selection index using additive genomic relationships (GSSI) and using a Gaussian kernel (KSSI). These models are described below; for simplicity, since all phenotypes were centered, we present models without intercept nor fixed effects. With the crossing and experimental design used, there were no common hybrids across years; therefore, the mixed-effects models did not include *genotype* × *year* interaction terms to account for its variability and relied only on genetic relationships among individuals.

GBLUP model. After adequate centering, the standard GBLUP model is represented by the following equation

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y} = [y_1,\ldots,y_n]'$, $\boldsymbol{u} = [u_1,\ldots,u_n]'$, and $\boldsymbol{\varepsilon} = [\varepsilon_1,\ldots,\varepsilon_n]'$ are t he vectors of adjusted phenotypes, breeding values (BV) and environmental error terms, respectively. Breeding values and errors are assumed to be normally distributed $\boldsymbol{u} \sim N(\mathbf{0}, \sigma^2_u\mathbf{G})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2_\varepsilon\mathbf{I})$, where σ^2_u and σ^2_ε are the genetic and error variances, respectively, $\mathbf{G}$ is the additive genetic relationship matrix (GRM), and $\mathbf{I}$ is an identity matrix.

The genomic relationship matrix $\mathbf{G} = \{g_{ij}\}$ was derived from markers, $\mathbf{X} = \{x_{im}\}$ using $\mathbf{G} = \mathbf{ZZ}'/p$, where $p = 4673$ is the number of markers and $\mathbf{Z} = \{(x_{im} - \bar{x}_m)/sd_{x_m}\}$ is the matrix of centered and scaled markers obtained by subtracting the mean of the corresponding column from each marker entry, followed by scaling by the standard deviation of the column.

The predicted BVs ($\widehat{\boldsymbol{u}}_{PS}$) for the individuals in the prediction set (PS) are then linear combinations of the phenotypes of the cultivars in the training set (TS, $\boldsymbol{y}_{TS}$) (Searle et al. 1992) such that

$$\widehat{\boldsymbol{u}}_{PS} = \mathbf{B}_G\boldsymbol{y}_{TS} \tag{2}$$

where $\mathbf{B}_G = \mathbf{G}_{PS,TS}(\mathbf{G}_{TS} + \lambda_0\mathbf{I})^{-1}$ is a Hat matrix (i.e., coefficients of regression of BVs on phenotypes), $\lambda_0 = \sigma^2_\varepsilon/\sigma^2_u$ is the ratio between residual and genetic variances, $\mathbf{G}_{PS,TS}$ is a matrix containing the additive genetic relationships between the data points in the prediction set and those in the training set, and $\mathbf{G}_{TS}$ represents the additive GRM of the training data points. The predictions of the GBLUP (given by Eq. (2)) can be shown to be equivalent to those of a selection index (e.g., Henderson 1977; Lopez-Cruz and de los Campos 2021).

RKHS models. RKHS regression is equivalent to the GBLUP model just described, with $\mathbf{G}$ replaced by any positive-definite kernel ($\mathbf{K}$). Here, we used a Gaussian kernel $\mathbf{K} = \{K_{ij}(\theta)\}$ $(i,j = 1,\ldots, n)$ where $K_{ij}(\theta) = \exp\left(-\theta\widetilde{d}^2_{ij}\right)$. Here θ is a bandwidth parameter and $\widetilde{d}^2_{ij}$ is the scaled squared Euclidean distance between individuals i and j given by their marker genotypes, obtained by dividing the distance $d^2_{ij} = \sum_{m=1}^{p}\left(x_{im} - x_{jm}\right)^2$ by the average distance $\overline{d} = \frac{1}{n^2}\sum_i\sum_j d^2_{ij}$. Following González-Camacho et al. (2012) we generated three different kernels $\mathbf{K}_1 = \{K_{ij}(\theta_1)\}$, $\mathbf{K}_2 = \{K_{ij}(\theta_2)\}$, and $\mathbf{K}_3 = \{K_{ij}(\theta_3)\}$, where $\theta_1 = 0.2$, $\theta_2 = 1$, and $\theta_3 = 5$.

In addition to the single-kernel models above-described, we also considered a multi-kernel model (aka, 'kernel averaging', KA; de los Campos et al. 2010) with the three kernels previously described. Briefly, The KA model includes three random effects, $\boldsymbol{y} = \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3 + \boldsymbol{\varepsilon}$, each following a normal distribution with its own variance parameter, $\boldsymbol{u}_k \sim N\left(\mathbf{0}, \sigma^2_{a_k}\mathbf{K}_k\right)$ $(k = 1,2,3)$.

Sparse selection index. This approach was recently introduced by Lopez-Cruz and de los Campos (2021). The methodology introduces an L1-sparsity-inducing penalty into a selection index problem. Here, in an SSI using additive relationships (GSSI), the weights of the selection index for the i^{th} individual in the prediction set was obtained from the following L1-penalized optimization problem

$$\widetilde{\boldsymbol{b}}_{i_G}(\lambda) = \arg\min_{\boldsymbol{b}_i}\left\{\frac{1}{2}\boldsymbol{b}_i'(\mathbf{G}_{TS} + \lambda_0\mathbf{I})\boldsymbol{b}_i - \mathbf{G}_i'\boldsymbol{b}_i + \lambda\sum_{j=1}^{n}\left|b_{ij}\right|\right\} \tag{3}$$

where $\mathbf{G}_i' = \mathbf{G}_{PS(i),TS}$ is the vector containing the additive relationships between the ith subject in the prediction set and each of the subjects in the training set, λ is a parameter controlling the degree of sparsity of $\widetilde{\boldsymbol{b}}_i$, and $\sum_{j=1}^{n}\left|b_{ij}\right|$ is an L1-penalty on the coefficients $\mathbf{b}_i$. A (sparse) Hat matrix for the GSSI, $\widetilde{\mathbf{B}}_G(\lambda)$ contains, in each row, the solutions to Eq. (3), obtained for each testing line, i.e., $\widetilde{\mathbf{B}}_G(\lambda) = \{\widetilde{\boldsymbol{b}}_{i_G}(\lambda)'\}$. A value of $\lambda = 0$ yields the same (non-sparse) Hat matrix as the standard GBLUP in Eq. (2).

For the SSI with a Gaussian kernel (KSSI), we used Eq. (3) with the kernel (either $\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{K}_3$, or $\mathbf{K}_A$) instead of the additive relationship matrix ($\mathbf{G}$) to obtain a sparse Hat matrix $\widetilde{\mathbf{B}}_K(\lambda) = \{\widetilde{\boldsymbol{b}}_{i_K}(\lambda)'\}$.

Although the optimization problem in Eq. (3) does not have a closed form, solutions can be derived using a coordinate descent algorithm (Lopez-Cruz et al. 2020). Finally, an optimal value of λ can be obtained using cross-validation within the training set.

Variance components

The implementation of GBLUP, KBLUP and the corresponding SSIs requires estimates of variance components. These estimates were obtained by fitting Bayesian genomic models into each trait-environment combination. These analyses were performed using the BGLR R-package (Perez and de los Campos 2014) with the default setting for hyper-parameters.

After fitting the models, posterior means of the variance components were obtained. For the standard KABLUP, the model was fitted with the three kernels together to estimate the kernel-specific variances and then used to derive $\sigma_a^2 = \sigma_{a_1}^2 + \sigma_{a_2}^2 + \sigma_{a_3}^2$ and the average kernel $\mathbf{K}_A = \frac{\sigma_{a_1}^2}{\sigma_a^2}\mathbf{K}_1 + \frac{\sigma_{a_2}^2}{\sigma_a^2}\mathbf{K}_2 + \frac{\sigma_{a_3}^2}{\sigma_a^2}\mathbf{K}_3$. The proportion of the trait variance that is explained by the BLUP models was calculated as $h^2 = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2)$ where σ_u^2 can be either the additive or non-additive (kernel) genetic variance estimate. All these models were fitted using only data from the training set.

Assessment of prediction accuracy

Variance components estimates and the corresponding GRM ($\mathbf{G}$, $\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{K}_3$, or $\mathbf{K}_A$) were used to derive the non-sparse ($\mathbf{B}_G$ or $\mathbf{B}_K$ for the standard BLUP) and the sparse ($\widetilde{\mathbf{B}}_G(\lambda)$ or $\widetilde{\mathbf{B}}_K(\lambda)$ for the SSI) Hat matrices. The predictions ($\widehat{\boldsymbol{u}}_{PS}$) were derived (as in Eq. (2)) as the product of the (non-sparse or sparse) Hat matrix times the vector of phenotypes in the training set. Prediction accuracy was measured as the correlation between observed and predicted values in the prediction set, i.e., $\rho = cor(\boldsymbol{y}_{PS}, \widehat{\boldsymbol{u}}_{PS})$.

The prediction accuracy was assessed for different prediction scenarios using cycle 2020 as the prediction set with different training set compositions using data from previous generations, as follows: (*i*) the data from the 2020 cycle was randomly partitioned into 85–15% (i.e., 413 and 73 individuals), (*ii*) the 85%-set ($n_{PS} = 413$) from the year 2020 was predicted using data of the single year 2017 ($n_{TS} = 901$), 2018 ($n_{TS} = 1418$), or 2019 ($n_{TS} = 722$), or cumulated years 2018+2019 ($n_{TS} = 2140$) or 2017 +2018+2019 ($n_{TS} = 3041$) as a training set, (*iii*) the prediction of the 413 individuals was also performed using the same training sets, but augmented by progressively including the remaining 15%-set from 2020, first 25 (5%), then 49 (10%), and finally, 73 (15%) individuals. See Table 1 for a summary of all the training set compositions. All predictions were performed 100 times using different random partitions of the 2020 data.

Software

All the aforementioned analyses were performed in the R environment-language (R Core Team 2019). All standard Bayesian GBLUP and KBLUP models were fitted using the BGLR R-package (Perez and de los Campos 2014) to estimate variance components. The sparse Hat matrices ($\widetilde{\mathbf{B}}_G(\lambda)$ or $\widetilde{\mathbf{B}}_K(\lambda)$) were obtained using the SFSI R-package (Lopez-Cruz et al. 2020). For each trait-environment partition, an optimal value of λ was obtained using 10-fold cross-validation within the training set.

RESULTS

The germplasm used in this study is derived from different biparental families across 4 years. This richness of the data is reflected in a high population heterogeneity, in which individuals cluster into groups within and across generations (Fig. 1). However, the crosses performed prevented the formation of a clear structure (e.g., 2 clusters); instead, the population shows a more cryptic substructure with varying degrees of admixture between families and generations. The intermixing between generations observed in Fig. 1 can be attributed to the connection among years through common parents leading to the formation of a varied number of half-sib families among all years.

Table 1. Training set (TS) composition used in each prediction scenario. (The prediction set was the same for all training scenarios and consisted of 413 (i.e., 85%) randomly chosen individuals from the 2020 cycle).

	% of 2020 data used for training (n)			
	0 (0)	5 (25)	10 (49)	15 (73)
TS cycle(s)	Total training size (n_{TS})			
2017	901	926	950	974
2018	1418	1443	1467	1491
2019	722	747	771	795
2018+2019	2140	2165	2189	2213
2017+2018+2019	3041	3066	3090	3114

Prediction accuracy comparison of GBLUP and KBLUP models

Figure 2 shows the accuracy of prediction (averaged across all 100 partitions) for GY in the optimal environment using all standard BLUP models for all different training set compositions representing a combination of previous cycles (2017, 2018, 2019, 2018–2019, or 2017–2019) plus the inclusion of 0, 5, 10, and 15% (i.e., 0, 25, 49, and 73 subjects) of the total number of individuals from the same prediction cycle 2020 (see Table 1). The lowest accuracies were observed when the training set was not referenced to the target prediction set (i.e., when no data from 2020 is included in the training set, compare the top panel with other panels in Fig. 2). As expected, the inclusion of individuals from the same prediction cycle increases the prediction accuracy across all models and training set composition. For instance, when augmenting the training set to include 25 genotypes (i.e., 5%) from 2020, the accuracy of the GBLUP increased by 88% when predicting with 2019 (0.18 vs 0.34) and by more than 100% when using 2018 (0.08 vs 0.32) and 2017 (0.02 vs 0.30) as a training set (compare the 2 top panels in Fig. 2). Furthermore, the increase in accuracy of the GBLUP is even larger (at least 180%) when adding 73 genotypes (i.e., 15%) from the 2020 cycle. The same patterns were obtained for GY-drought (Supplementary Fig. S1) in which the accuracy of the models is very low when no reference data from 2020 is included in the training set and it increases as reference data is added to the training set.

Higher prediction accuracies are achieved when using only the most recent previous generation (cycle 2019) as a training set (see the top panel in Fig. 2); this was expected as the number of parents shared between the prediction cycle 2020 and other cycles (Fig. 1C) was larger for cycle 2019 (11 parents) than with previous cycles (3 parents shared between cycle 2020 and either cycle 2018 or 2017). However, older generations (2018 or 2017), when combined with data from 2020, provided similar or even better predictions. For instance, the BLUP models trained with 2017 (plus 15% of the 2020 cycle) data yielded accuracies that are 0.02–0.04 points (in the correlation scale) higher than that of models trained with 2019 plus 15% of the 2020 data (see the bottom panel in Fig. 2).

The results when cumulated previous years (2018+2019 and 2017+2018+2019) were used as a training set to predict 2020 are also shown in Fig. 2. The highest prediction accuracies were obtained when using one generation back (i.e., cycle 2019 alone) as a training set. The inclusion of older generations in the training set did not increase but rather reduced the accuracy, compared with using only one generation back (Fig. 2). For instance, using 2018–2019 as a training set, the accuracy of the GBLUP model showed a reduction of 5–6% (0.01–0.03 points) relative to when using 2019 alone. Using three generations back (2017–2019) for model training yielded a decrease of 15–30% (0.05–0.08 points) in the accuracy. The same patterns were also observed for GY-drought (Figure S1) in which the accuracy of the models was reduced by 12–60% (0.04–0.17 points, relative to when using 2019 alone as a training set) as more older generations were cumulated in the training set.

In general, kernel-based models (especially using the Gaussian kernels $\mathbf{K}_1$, $\mathbf{K}_2$, and $\mathbf{K}_A$) achieved higher prediction accuracy than the standard GBLUP with increases in accuracy of 1–15% (0.01–0.06 points). Although kernels $\mathbf{K}_1$, $\mathbf{K}_2$, and $\mathbf{K}_3$ are ranked differently across training set compositions, models with $\mathbf{K}_A$ seem to be more stable across all scenarios, with a performance similar to that of the best of the three kernels $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_3$. These results are in agreement with the findings of de los Campos et al. (2010) who highlighted the importance of combining different kernels as a way to make the model robust with respect to the choice of kernel.

Effect of sparsity on prediction accuracy

The same partitions of training-prediction sets used to obtain the results for the (non-sparse) BLUP models were used to evaluate

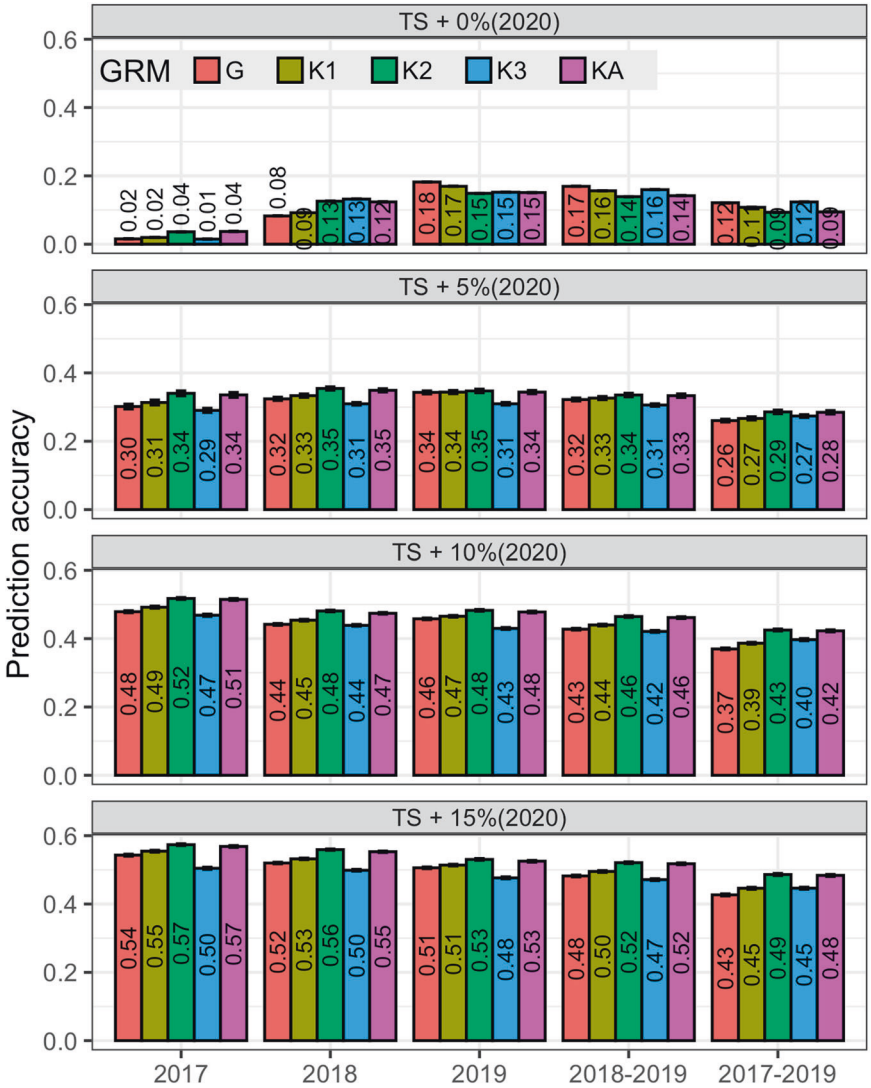

Fig. 2 Prediction accuracy of the BLUP models by training set (TS). Models were fitted using different genetic relationship matrices (**G**, $\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{K}_3$, or $\mathbf{K}_A$). TSs consisted of all the data from the single cycles 2019, 2018, or 2017 alone (top-left panel), or in combination with a proportion (5% = 25, 10% = 49, 15% = 73) of the data from the 2020 cycle. The prediction set consisted of 413 genotypes (representing the 85%) from the 2020 cycle that were not used for model training. Trait GY, optimal environment.

the prediction accuracy of SSIs (sparse models). A cross-validated value λ_{CV} was found within the training set to calculate an optimal SSI. Table 2 contains the results of the predictions of GY-optimal trait-environment combination for the scenario in which 15% of the data from 2020 is included in the training set (either 2017, 2018, 2019, 2018–2019, or 2017–2019). Results for the cases when adding 0, 5, and 10% of the 2020 data are presented in Supplementary Tables S1–S3. With this training set composition, the accuracy of the standard GBLUP models was between 0.43 and 0.54.

Standard KBLUP (with kernels $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_A$) models achieved higher prediction accuracy than the standard GBLUP, with gains in accuracy (relative to the GBLUP) ranging from minimal (0.01 points) to substantial (0.06 points). Sparse models (GSSI and KSSI) yielded even higher accuracies than the standard GBLUP, with gains in accuracy (relative to the GBLUP) ranging from 0.02 to 0.08 points (Table 2). The gains in prediction accuracy of the KBLUP models are smaller when the accuracy of the standard GBLUP models was very low. For instance, when the accuracy of the GBLUP is as low as 0.02–0.18 (the case where no data from 2020 are included in the training set, Supplementary Table S1), standard KBLUP models performed similarly or worse (0.01–0.03 reduction in accuracy) than the standard GBLUP models; however, SSIs yielded gains in accuracy (relative to the standard GBLUP) of 0.01–0.08 (Supplementary Table S1). It was only in these low-accuracy situations that, in some cases, the use of a standard KBLUP model resulted in ~0.03 loss of accuracy (relative to the standard GBLUP), and that using sparse models, the accuracy lost was 0.01–0.05 (see Supplementary Table S1).

In the situations where data from 2020 was used for model training, the sparse models (GSSI or KSSI) provided always an increase in the accuracy, relative to their corresponding non-sparse (GBLUP or KBLUP) models, of 0.02–0.08 points (Table 2 and Supplementary Tables S2 and S3); this was true only for models using additive relationships **G** and kernels $\mathbf{K}_1$, $\mathbf{K}_2$, or $\mathbf{K}_A$. However, when no data from 2020 is included in the training set (i.e., the low-accuracy cases), the SSIs showed a reduction (relative to the standard BLUPs) in the accuracy of 0.03–0.05 points (Supplementary Table S1). No significant difference between sparse and non-sparse models was observed when using the large-bandwidth kernel $\mathbf{K}_3$.

In summary, across all scenarios, the standard GBLUP showed the lowest accuracy among all models (except the $\mathbf{K}_3$-based

Table 2. Accuracy of prediction for each training set (TS) composition (including 15% = 73 subjects from the 2020 cycle), trait GY, optimal environment.

TS/n_{TS}	GRM	λ_{CV}[a]	n_{sup}[b] (%sparsity)	h^2	Accuracy (SD)		Gain 1 (%)[c]	Gain 2 (%)[d]
					BLUP	SSI		
2017 n_{TS} = 974	G	0.0175	151 (16)	0.53	0.54 (0.079)	0.56 (0.079)	–	3
	K1	0.0037	180 (18)	0.87	0.55 (0.077)	0.54 (0.089)	2	−3
	K2	0.0048	238 (24)	0.76	0.57 (0.071)	0.57 (0.072)	6	0
	KA	0.0030	293 (30)	0.85	0.57 (0.074)	0.57 (0.077)	5	0
2018 n_{TS} = 1491	G	0.0089	358 (24)	0.61	0.52 (0.064)	0.57 (0.052)	–	10
	K1	0.0017	414 (28)	0.91	0.53 (0.063)	0.57 (0.053)	2	8
	K2	0.0028	508 (34)	0.80	0.56 (0.057)	0.58 (0.053)	8	4
	KA	0.0018	498 (33)	0.88	0.55 (0.058)	0.58 (0.053)	6	4
2019 n_{TS} = 795	G	0.0357	71 (9)	0.54	0.51 (0.068)	0.54 (0.066)	–	7
	K1	0.0139	90 (11)	0.88	0.51 (0.068)	0.55 (0.066)	2	7
	K2	0.0058	207 (26)	0.73	0.53 (0.069)	0.54 (0.068)	5	2
	KA	0.0041	219 (28)	0.80	0.53 (0.069)	0.54 (0.070)	4	2
2018–2019 n_{TS} = 2213	G	0.0148	246 (11)	0.57	0.48 (0.071)	0.54 (0.063)	–	12
	K1	0.0024	373 (17)	0.90	0.50 (0.071)	0.54 (0.064)	3	8
	K2	0.0023	783 (35)	0.77	0.52 (0.070)	0.54 (0.066)	8	3
	KA	0.0016	864 (39)	0.82	0.52 (0.071)	0.54 (0.066)	7	3
2017–2019 n_{TS} = 3114	G	0.0141	322 (10)	0.52	0.43 (0.079)	0.49 (0.081)	–	14
	K1	0.0026	435 (14)	0.88	0.45 (0.081)	0.51 (0.072)	5	13
	K2	0.0023	978 (31)	0.75	0.49 (0.081)	0.51 (0.079)	14	4
	KA	0.0018	1079 (35)	0.81	0.48 (0.081)	0.50 (0.078)	13	4

GRM genetic relationship matrix, *SD* standard deviation, h^2 proportion of the trait variance explained by the model.
[a]Penalization parameter in Eq. (3) found by cross-validating the TS.
[b]n_{sup} = average number of individuals from the TS with a non-zero coefficient in the sparse Hat matrix (support set). %sparsity = $100 \times n_{TS}/n_{sup}$. In the BLUP models, λ_{CV} is equal to zero and n_{sup} is equal to the total TS size. Within each TS cycle, percentage of increase in accuracy of [c]the standard KBLUP relative to the standard GBLUP ($= 100 \times \frac{KBLUP\ accuracy - GBLUP\ accuracy}{GBLUP\ accuracy}$), and of [d]the *SSI relative to the standard *BLUP ($= 100 \times \frac{*SSI\ accuracy - *BLUP\ accuracy}{*BLUP\ accuracy}$, * = G, K_1, K_2, or K_A).

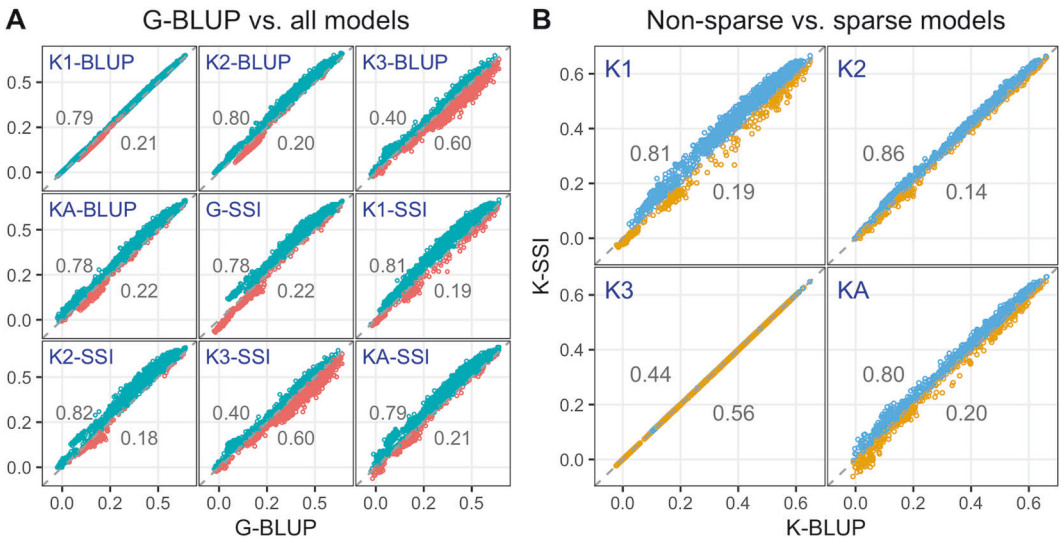

Fig. 3 Point-wise prediction accuracy comparison between models. A Prediction accuracy of the standard (non-sparse) GBLUP model (horizontal axis) versus the prediction accuracy of all other models (vertical axis of each panel). **B** Prediction accuracy of the standard KBLUP model (horizontal axis) versus the prediction accuracy of the KSSI (vertical axis) by type of kernel used in panels. Each point represents a training-testing partition within each training set composition. Colored dots (numbers) above (below) the 45-degree line represent cases (the proportion) for which one model outperformed the other model. Trait GY, optimal environment.

models, see Fig. 3A). This inferiority of the standard GBLUP was also observed for GY in the drought environment (see Supplementary Fig. S2A). The addition of sparsity to the KBLUP models resulted sometimes (in at least 80% of the cases) in an increased accuracy when the kernel used a small bandwidth ($\mathbf{K}_1$ with $\theta = 0.2$, and $\mathbf{K}_2$ with $\theta = 1$) or when averaged across extreme kernels ($\mathbf{K}_A$) for optimal and drought environments (Fig. 3B and Supplementary Fig. S2B).

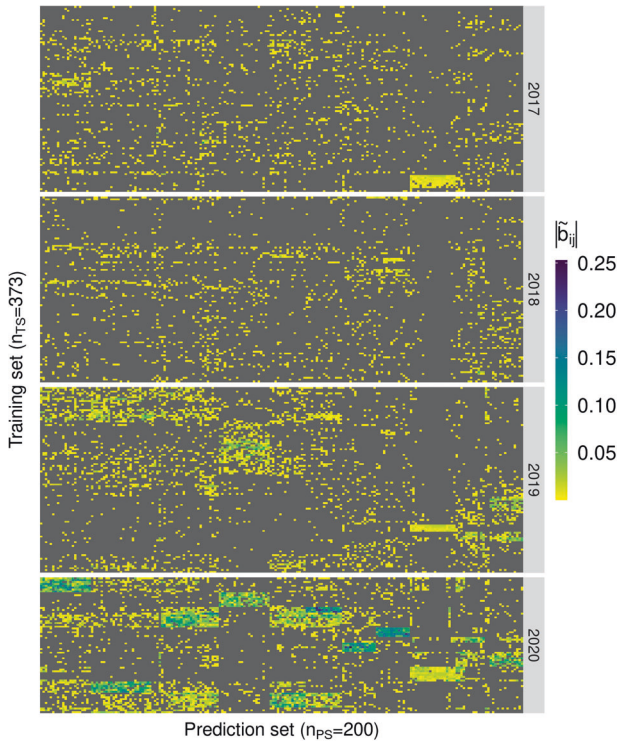

Fig. 4 Heatmap of the coefficients in the Hat matrix ($\widetilde{\mathbf{B}}_G(\lambda)$) of the GSSI for one training-prediction (TS-PS) partition in the prediction of $n_{PS} = 413$ individuals from 2020 using $n_{TS} = 3114$ individuals (2017+2018+2019 plus 15% = 73 genotypes from the 2020 cycle). Columns represent (a sample of 200) predicted individuals and rows represent (a sample of 100 individuals from each cycle 2017–2019 and the 73 subjects from 2020) training individuals, separated by cycle. Each column vector represents values of the vector $\widetilde{\boldsymbol{b}}_{i_G}(\lambda) = \{\widetilde{b}_{ij}\}$, $j = 1,\ldots,3114$ (Eq. (3)) using a value of λ obtained by cross-validation. Individuals not contributing to the prediction have a coefficient $\widetilde{b}_{ij} = 0$ and are in gray. Individuals with a non-zero coefficient are shown in a yellow-blue scale. Trait GY, optimal environment.

Automatic training-sample selection

Table 2 (and Supplementary Tables S1–S3) show the optimal value of the penalization parameter λ and the degree of sparsity of the resulting SSI, measured by the average number of subjects from the training set in the support set (n_{sup}, subjects with a non-zero coefficient in the estimated Hat matrix) of each predicted genotype. The degree of sparsity varied across models. For GY in optimal environments, across all training set compositions, the strongest sparsity was achieved using the genomic matrix **G** with a relative sparsity (n_{sup}/n_{TS}) of 4–42% (Table 2 and Supplementary Tables S1–S3), while the relative sparsity of the KSSIs increases as the bandwidth parameter θ increases (relative sparsity of 9–52% for $\mathbf{K}_1$ and 20–62% for $\mathbf{K}_2$). The relative sparsity achieved when using the $\mathbf{K}_A$ kernel (21–59%) was similar to that of the $\mathbf{K}_2$ kernel (see Table 2 and Supplementary Tables S1–S3). The fact that no differences in accuracy were observed between the standard K_3BLUP and K_3SSI is due to the fact that the optimal λ_{CV} was zero; thus, the sparse model was equivalent to the standard model.

Figure 4 displays a heatmap of the sparse Hat matrix ($\widetilde{\mathbf{B}}_G(\lambda)$) of the GSSI. Selected individuals in the training set (2017–2019 plus 15% of data from 2020) appear in rows and those in the prediction set are shown in columns. Individuals from the training set that did not contribute to each SSI (i.e., those with zero weight in the index) are displayed in gray. Those with a non-zero coefficient (support set) are shown in a yellow-blue scale. The heatmap shows how SSIs select custom training sets for each genotype in the prediction set. Individual genotypes in the training set support the prediction of some, though not all the genotypes in the prediction set. The solution for the Hat matrix in Fig. 4 is very sparse, with a varying number of support points (comprising data from all generations) by testing genotype. Training genotypes from the same prediction cycle 2020 are more important for prediction (as measured by the magnitude of the regression coefficient) as they are more closely related to testing individuals. These individuals had the higher coefficients in the Hat matrix of the GBLUP and were not zeroed-out in the GSSI (see Supplementary Fig. S3). Prediction of each of the 413 testing genotypes was performed using phenotypes from an average of 322 (out of 3114, i.e., 10%; see Table 2) training genotypes. For the same prediction scenario, a heatmap for the KSSI with kernel $\mathbf{K}_A$ (showing 35% sparsity) is provided in Supplementary Fig. S4.

Figure 5 shows, for each of the sparse models, the proportion of the training individuals from each cycle (2017, 2018, or 2019) that contributed to the prediction (within the cycle support set) of the testing individuals. Each panel represents the different training sets composed of 2017+2018+2019 data plus the addition of either 0, 5, 10, or 15% of the 2020 data. As expected, training individuals that belong to the same group as the testing individuals are more likely to be included in the support set. For example, using a GSSI trained with 2017–2019 plus 5% from the 2020 data (see the top-right panel in Fig. 5), on average, 22% of the 25 genotypes from 2020 included in the training set contributed to the prediction of the testing individuals. Although more abundant, a smaller portion of the total individuals from previous cycles (11% of the 901 subjects from 2017 cycle, 9% of the 1418 from 2018, and 15% of the 722 from 2019) also contributed to the prediction. With a smaller degree of sparsity, similar patterns were also observed for the KSSIs (Fig. 5) except with $\mathbf{K}_3$, which rendered no sparsity (not shown in the figure). Plots displaying the within-cycle sparsity patterns for GY in the drought environment are shown in Supplementary Fig. S5.

As more individuals from the same prediction cycle are added to the training set, fewer individuals from previous generations become less frequent in the support set. For instance, performing the prediction with a GSSI using 2017–2019 cycles including 15% of the 2020 data (bottom-right panel in Fig. 5), yielded a smaller support set with only 10% (90 of 901) of the 2017 data, 8% (114 of 1418) of the 2018 data, and 14% (101 of 722) of the data from 2019; however, data from 2020 became more frequent in the support set with 24% (18 of 73 subjects). This happened because the SSI zeroed-out the regression coefficient (b_{ij}) of training individuals that, in the Hat matrix of the BLUP (Eq. (2)), are already relatively small in absolute value and that have a small absolute genomic relationship (g_{ij}) with individuals in the prediction set (i.e., training individuals in a neighborhood of the origin); these coefficients correspond mostly to individuals from previous cycles (see Supplementary Fig. S3).

DISCUSSION

Multiple factors can affect the predictive performance of GS models, including sample size, trait heritability, the extent of LD between markers and quantitative trait loci (QTL), as well as the relationships between training and testing genotypes (Daetwyler et al. 2008; Heffner et al. 2009; Lorenzana and Bernardo 2009; Combs and Bernardo 2013).

General guidelines suggest that prediction accuracy is maximized when the training set includes a sufficiently large number of individuals distantly related to each other (Rincent et al. 2012) and closely related to the subjects in the prediction set (Habier et al. 2010; Clark et al. 2012). On the other hand, there is evidence suggesting that increasing the training set size by including

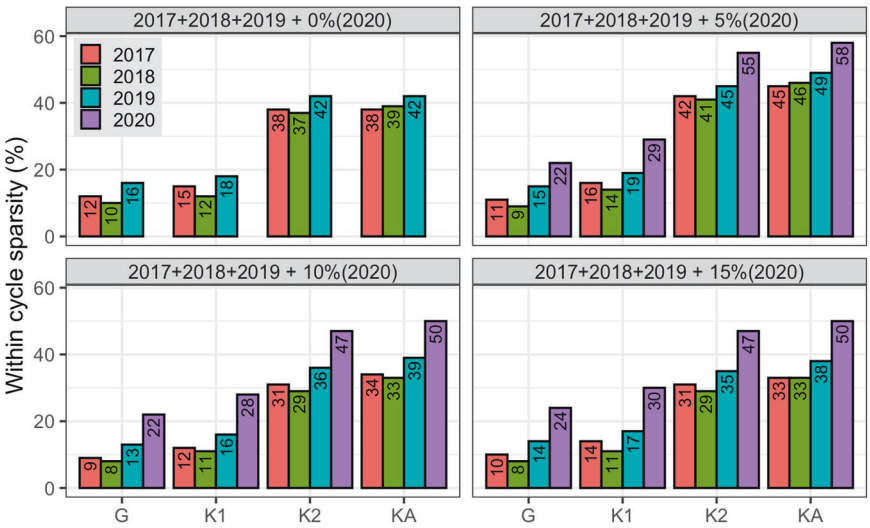

Fig. 5 Proportion of the training individuals from each cycle that contributed to the prediction of genotypes from 2020 (averaged across all the 413 testing subjects), using SSIs with different relationship matrices (G, K_1, K_2, or K_A). The training set was composed of individuals from 2017 ($n = 901$), 2018 ($n = 1418$), and 2019 ($n = 722$) alone (top-left panel), or in combination with a proportion (5% = 25, 10% = 49, 15% = 73) of the data from the 2020 cycle. Trait GY, optimal environment.

individuals that are genetically distant to those in the prediction set does not necessarily increase, and might even reduce, the prediction accuracy (e.g., Lorenz and Smith 2015).

Each cycle of a breeding program produces a new batch of genotype/phenotype data; therefore, after many years of adopting GS, the data available for model training are typically multi-generational and may often include complex patterns of pedigree relationships within and between generations. There is clear evidence that a GS model needs to be re-trained every cycle (Wolc et al. 2011; Wientjes et al. 2013; Pszczola and Calus 2016). When re-training models, breeding organizations face many challenges. Should all the available data be used for model training? Should researchers restrict the training data to only include genotypes/phenotypes from recent generations? Or should they exclude data from genotypes distantly related to the current set of selection candidates?

Some evidence suggests that in genomic prediction, 'bigger is not necessarily better'. For instance, using historic wheat data generated over 17 years, Dawson et al. (2013) observed that the accuracy of year-to-year predictions using training sets composed of all previous years was approximately the same as when considering only three years back. Likewise, in a broiler breeding population, Wolc et al. (2016) found that the maximum accuracy was accomplished when the training set was composed of the three most recent generations.

The SSI methodology offers a framework to identify a customized training set (or support points) for each individual in the prediction set, from which the predictions are derived. This methodology considers both the relationships between the candidate for selection and each training genotype, as well as relationships between training genotypes (Eq. (3)). Therefore, we suggest that the SSI can be used to address the problem of training set optimization with multi-generation data. In this research we used multi-generation data originated from more than 50 biparental families to measure the impact of sparsity using SSIs formed using additive and non-additive kernels. We corroborated that the use of all available multi-generation information together can decrease the prediction accuracy of the standard GBLUP as the data become increasingly heterogeneous (see Fig. 2 and Supplementary Fig. S1) where the performance of genotypes is also affected by the usual variability in weather conditions from one year to another (not applicable in this study).

The accuracy of markers-derived predictions is provided by family relationships and by the LD between QTL and markers. These two sources of accuracy are difficult to separate (Habier et al. 2007; Habier et al. 2010). Increasing marker density can increase the proportion of the trait variance explained by BLUP models; however, closing the gap between the trait- and SNP-heritability requires having markers in high LD with causal variants (Makowsky et al. 2011; Kim et al. 2017). We observed that, when augmenting the training set by adding only a few individuals from the same cycle, the accuracy increased but the proportion of variance explained by the model remained unchanged (compare Table 2 and Supplementary Tables S1–S3); suggesting a sizable contribution of closely related individuals (i.e., individuals from the same cycle) to prediction accuracy (see Supplementary Fig. S3). These results are in agreement with Habier et al. (2007) who found that the accuracy of BLUP models is mostly resulted from the genetic relationships between training and testing individuals, with a bare contribution from LD.

One could expect that models that have higher proportion of variance explained in the training set (estimated using a ratio of variance components) may also achieve a higher prediction accuracy. However, this is not necessarily the case as there may be over-fitting or because prediction accuracy also depends on the accuracy of estimated effects (e.g., Goddard 2009; Makowsky et al. 2011). In general, the RKHS models fitted the training data better (these models had 0.20–0.35 points higher 'heritability' than the GBLUP, see Table 2). However, the difference in prediction accuracy between the linear and non-linear models was more modest; thus, suggesting that RKHS models either over-fitted the data or were more complex and, therefore, with the same training set, achieved a lower accuracy of estimates of effects.

Our results confirmed that an SSI based on additive relationships yields a higher prediction accuracy than the standard additive GBLUP. When a non-additive kernel was used, we found that sparsity improved prediction accuracy, provided that the kernel used was not already a 'local' kernel, that is, a kernel in which genetic covariances are positive only for closely related individuals. For local kernels ($\mathbf{K}_3$), adding sparsity did not improve prediction accuracy in a clear and systematic manner. Therefore, our results confirm the benefit of 'local predictions' (largely dependent on closely related individuals), which can be obtained either by using an RKHS with local kernels or with an SSI applied to additive genomic relationships.

Both the SSI and the Kernels regression require the optimization of a parameter that controls local predictions. The SSI requires the optimization of the penalization parameter (λ), which can be done by cross-validation within the training data set. On the other hand, the RKHS regression requires the bandwidth parameter, which controls how fast covariances drop with genetic distance, to be tuned. This can be done either by comparing multiple kernels using cross-validation or by using multiple kernels with 'kernel averaging,' as discussed in de los Campos et al. (2010).

Standard training-optimization methods assume that a single training set is optimal for all selection candidates. The SSI does not make this assumption. Our results clearly show that each SSI picks a particular set of support points and that the optimal training set varies between lines in the prediction set. The inspection of the Hat matrix of the SSI (see Fig. 4 and Supplementary Fig. S4) makes it clear that in prediction, *one-size-does-not-fit-all* selection candidates. Likewise, the inspection of the Hat matrix shows that optimizing training sets by restricting the training data to recent generations may also not be optimal. Indeed, most of the SSIs picked information from all the generations available, with varying levels of sparsity. Lopez-Cruz and de los Campos (2021) found that the SSI provided higher gains in accuracy (over the GBLUP) as the number of training subjects increased (keeping marker density and population structure fixed). Their results demonstrated that the SSI can be more advantageous over BLUP models in situations when data exhibits a high degree of heterogeneity and/or is very large. Under these conditions, the SSI has more opportunities to detect an optimal training set providing increased accuracy.

Computational considerations

For a given training and testing set, computing an SSI for n_{PS} testing genotypes involves solving a penalized problem (in Eq. (3)) n_{PS} times. This is generally computationally more costly than solving BLUP equations (Eq. (2)) for the same training and testing sets. However, it is worth noting that the task of solving SSIs for n_{PS} genotypes can be fully parallelized in n_{PS} independent tasks. The computational cost of solving the SSI for one testing genotype depends on the training size and on the optimal value of the regularization parameter λ (highly sparse indices are very fast to compute). In a Linux-based system with a 2.4 GHz Intel Xeon Gold processor and 32 GB of RAM, solving the penalized problem for a single testing individual using all three previous years plus 15% of 2020 data ($n_{TS} = 3114$) as training took on average 0.11 s for the GSSI and 2.01 s for the K_ASSI (see Supplementary Fig. S6). As noted, all these computations can be fully parallelized in a high-performance computing environment. For instance, calculating a GSSI for all the $n_{PS} = 413$ testing individuals can take around $413 \times \frac{0.11}{10} = 4.5$ seconds with parallel computing using 10 cores, while it can take up to $413 \times \frac{2.01}{10} = 83.0$ seconds to compute the K_ASSI. The function 'SSI' from the SFSI R-package offers a functionality for multi-core computing.

Conclusion

SSIs can be used to optimize prediction accuracy when the training data exhibit complex relationship patterns. In this context, differences in allele frequencies and in LD patterns may make SNP effects heterogenous across families and sub-families, thus making the standard GBLUP sub-optimal. Both local kernels and SSIs can be used to optimize prediction accuracy in such data sets.

DATA AVAILABILITY

Data used in this study is available through the Dryad data repository (https://doi.org/10.5061/dryad.qjq2bvqgz). Scripts showing how to perform all the analyses are contained in File S1. All supplemental figures and tables are provided in File S2.

REFERENCES

Akdemir D, Isidro-Sanchez J (2019) Design of training populations for selective phenotyping in genomic prediction. Sci Rep 9:1–15

Alvarado G, Rodríguez FM, Pacheco A, Burgueño J, Crossa J, Vargas M et al. (2020) META-R: A software to analyze data from multi-environment plant breeding trials. Crop J 8:745–756

Atanda SA, Olsen M, Burgueño J, Crossa J, Dzidzienyo D, Beyene Y et al. (2021) Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. Theor Appl Genet 134:279–294

Bandeira e Sousa M, Cuevas J, de Oliveira Couto EG, Perez-Rodriguez P, Jarquín D, Fritsche-Neto R et al. (2017) Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. G3 Genes Genomes Genet 7:1995–2014

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Beyene Y, Gowda M, Olsen M, Robbins KR, Pérez-Rodríguez P, Alvarado G et al. (2019) Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. Front Plant Sci 10:1–11

Buckler E, Ilut DC, Wang X, Kretzschmar T, Gore M, Mitchell SE (2016) rAmpSeq: Using repetitive sequences for robust genotyping. BioRxiv (Preprint)

Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44:1–9

Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:1–7

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Perez-Rodriguez P, de los Campos G (2017) Bayesian genomic prediction with genotype × environment interaction kernel models. G3 Genes Genomes Genet 7:41–53

Cuevas J, Crossa J, Soberanis V, Perez-Elizalde S, Perez-Rodriguez P, de los Campos G et al. (2016) Genomic prediction of genotype × environment interaction kernel regression models. Plant Genome J 9:1–20

Cuevas J, Granato I, Fritsche-Neto R, Montesinos-López OA, Burgueño J, Bandeira e Sousa M et al. (2018) Genomic-enabled prediction Kernel models with random intercepts for multi-environment trials. G3 Genes Genomes Genet 8:1347–1365

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:1–8

Dawson JC, Endelman JB, Heslot N, Crossa J, Poland J, Dreisigacker S et al. (2013) The use of unbalanced historical data for genomic selection in an international wheat breeding program. F Crop Res 154:12–22

de los Campos G, Gianola D, Rosa GJ (2009) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87:1883–1887

de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res 92:295–308

Garrick DJ (2011) The nature, scope and impact of genomic prediction in beef cattle in the United States. Genet Sel Evol 43:1–11

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776

Goddard M (2009) Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257

González-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G et al. (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:1–12

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51

Hazel LN (1943) The genetic basis for constructing selection indexes. Genetics 28:476–490

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Henderson CR (1977) Best linear unbiased prediction of breeding values not in the model for records. J Dairy Sci 60:783–787

Howard R, Gianola D, Montesinos-López O, Juliana P, Singh R, Poland J et al. (2019) Joint use of genome, pedigree, and their interaction with environment for

predicting the performance of wheat lines in new environments. G3 Genes Genomes Genet 9:2925–2934
Jacobson A, Lian L, Zhong S, Bernardo R (2014) General combining ability model for genomewide selection in a biparental cross. Crop Sci 54:895–905
Kim H, Grueneberg A, Vazquez AI, Hsu S, De Los Campos G (2017) Will big data close the missing heritability gap? Genetics 207:1135–1145
Lehermeier C, Schön CC, de los Campos G (2015) Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics 201:323–337
Lopez-Cruz M, de los Campos G (2021) Optimal breeding-value prediction using a Sparse Selection Index. Genetics 218:1–10
Lopez-Cruz M, Olson E, Rovere G, Crossa J, Dreisigacker S, Suchismita M et al. (2020) Regularized selection indices for breeding value prediction using hyper-spectral image data. Sci Rep 10:8195. https://doi.org/10.1038/s41598-020-65011-2
Lorenz AJ, Smith KP (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in Barley. Crop Sci 55:2657–2667
Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161
Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7:1–9
Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
Morota G, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. Front Genet 5:1–13
Olson KM, VanRaden PM, Tooker ME (2012) Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci 95:5378–5383
Perez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495
Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y et al. (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome J 5:103–113
Pszczola M, Calus MPL (2016) Updating the reference population to achieve constant genomic prediction reliability across generations. Animal 10:1018–1024
R Core Team (2019) R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194:493–503
Rincent R, Nicolas S, Altmann T, Brunel D, Revilla P, Melchinger A et al. (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192:715–728
Rio S, Moreau L, Charcosset A, Mary-Huard T (2020) Accounting for group-specific allele effects and admixture in genomic predictions: theory and experimental evaluation in maize. Genetics 216:27–41
Roth M, Muranty H, Di Guardo M, Guerra W, Patocchi A, Costa F (2020) Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. Hortic Res 7:148. https://doi.org/10.1038/s41438-020-00370-5
Searle SR, Casella G, McCulloch CE (1992) Variance components. John Wiley & Sons, Inc. Hoboken, New Jersey
Smith HF (1936) A discrimant function for plant selection. Ann Eugen 7:240–250
VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423
Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–631
Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R et al. (2011) Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet Sel Evol 43:1–8
Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP et al. (2016) Implementation of genomic selection in the poultry industry. Anim Front 6:23–31
Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C et al. (2020) Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun 1:1–21

ACKNOWLEDGEMENTS

Field experiments were funded by the Bill and Melinda Gates Foundation, and the United States Agency for International Development (USAID) through the Stress Tolerant Maize for Africa (STMA, #OPP1134248) and the CGIAR Research Program MAIZE. The CGIAR Research Program MAIZE receives W1&W2 support from the Governments of Australia, Belgium, Canada, China, France, India, Japan, Korea, Mexico, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom, United States, and the World Bank. The lead author acknowledges CIMMYT Global Maize Program, under CGIAR Research Program MAIZE and the Robbins Lab at Cornell University that provided field and marker data for this study. MLC was supported by the Monsanto's Beachell-Borlaug International Scholarship Program (MBBISP) and by the Dissertation Completion Fellowship funded by the Michigan State University Graduate School. GDLC received support from the National Institute for Food and Agriculture (NIFA) of the USDA (award #2021-67015-33413).

AUTHOR CONTRIBUTIONS

MLC, JC, and GDLC conceptualized the study; MLC and PPR performed the statistical analyses. YB and MG performed field experiments. GDLC, JC, YB, and MG acquired funding sources. All authors contributed equally on writing the manuscript.

Compliance with ethical standards

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/s41437-021-00474-1.

Correspondence and requests for materials should be addressed to Marco Lopez-Cruz.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.