Theoretical and Applied Genetics manuscript No.
(will be inserted by the editor)

Genomic prediction and training set optimization in a structured Mediterranean oat population

Simon Rio* · Luis Gallego-Sánchez* · Gracia Montilla-Bascón · Francisco J. Canales · Julio Isidro y Sánchez · Elena Prats

Received: date / Accepted: date
∗ These authors contributed equally to this study

Abstract Key message: The strong genetic structure observed in Mediterranean oats affects the predictive ability of genomic prediction as well as the performance of training set optimization methods. In this study, we investigated the efficiency of genomic prediction and training set optimization in a highly structured population of cultivars and landraces of cultivated oat (*Avena sativa*) from the Mediterranean basin, including white (subsp. *sativa*) and red (subsp. *byzantina*) oats, genotyped using genotype-by-sequencing markers, and evaluated for agronomic traits in Southern Spain. For most traits, the predictive abilities were moderate to high with little differences between models, except for biomass for which Bayes-B showed a substantial gain compared to other models. The consistency between the structure of the training population and the population to be predicted was key to the predictive ability of genomic predictions. The predictive ability of inter-subspecies predictions was indeed much lower than that of intra-subspecies predictions for all traits. Regarding training set optimization, the linear mixed model optimization criteria (PEVmean and CDmean) performed better than the heuristic approach "partitioning around medoids", even under high population structure. The superiority of CDmean and PEVmean could be explained by their ability to adapt the representation of each genetic group according to those represented in the population to be predicted. These results represent an important step towards the implementation of genomic prediction in oat breeding programs and address important issues faced by the genomic prediction community regarding population structure and training set optimization.

Simon Rio · Julio Isidro y Sánchez
Centro de Biotecnologia y Genómica de Plantas (CBGP, UPM-INIA) Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnologia Agraria y Alimentaria (INIA) Campus de Montegancedo-UPM 28223-Pozuelo de Alarcón, (Madrid), Spain
E-mail: j.isidro@upm.es

Luis Gallego-Sánchez · Gracia Montilla-Bascón · Francisco J. Canales · Elena Prats
Institute for Sustainable Agriculture, Spanish Research Council (CSIC), Córdoba, Spain
E-mail: elena.prats@ias.csic.es

Keywords Oat · Avena sativa · Genomic prediction · Training set optimization · Genetic structure · Environmental adaptation

Introduction

The cultivated oat (*Avena sativa* L.) is an important and versatile crop that can be grown for grain, forage, feed and straw (Welch, 2012). In the last twenty years, the oat cultivation area within the Mediterranean region increased with an approximate rate of 7,500 Ha per year (FAO., 2017). However, the yields achieved in Northern Europe remains much greater than that of the Mediterranean area, as a result of the limited adaptation of oats to the environmental conditions of Southern Europe (e.g., water availability or temperature). Thus, there is a critical need for selecting oat genotypes that are better adapted to the Mediterranean environment by taking advantage of the existing diversity in the region: the white (subsp. *sativa*) and red (subsp. *byzantina*) oats.

The advent of molecular markers has revolutionized the methodologies for selecting individuals in animal and plant breeding programs. Among these new methods, genomic prediction (GP) was proposed by Meuwissen et al. (2001) and stands today as one of the most promising tool. In its simplest application, a set of individuals is evaluated for a given trait and genotyped using single nucleotide polymorphisms (SNPs). A statistical model is trained on this data set, referred to as the training set (TRS), and is used to predict the breeding value of individuals for whom only SNP information is available, referred to as the test set (TS) (Isidro et al., 2016). Several methods have been proposed in the literature including models making different hypotheses on the distribution of the effects of quantitative trait loci (QTL) like GBLUP, BayesA, BayesB or BayesCπ (Meuwissen et al., 2001; Habier et al., 2011; Heslot et al., 2012), semi-parametric methods like the reproducing kernel Hilbert space (RKHS) (Gianola et al., 2006; Gianola and van Kaam, 2008), or tree-based methods like random forests (Breiman, 2001; Chen and Ishwaran, 2012).

One of the most critical steps in GP is the selection of the TRS since it is critical to the predictive ability of the models. In the last few years, several studies have investigated different approaches to optimize TRSs (Rincent et al., 2012; Hickey et al., 2014; Akdemir et al., 2015; Isidro et al., 2015; Lorenz and Smith, 2015; Tayeh et al., 2015; Akdemir, 2017; Rincent et al., 2017; Brandariz and Bernardo, 2018; Norman et al., 2018; Akdemir and Isidro-Sánchez, 2019; Berro et al., 2019; Edwards et al., 2019; Guo et al., 2019; Mangin et al., 2019; Ou and Liao, 2019; Sarinelli et al., 2019; Alvarenga et al., 2020; Olatoye et al., 2020; Roth et al., 2020). Among the optimization criteria, some approaches have been the subject of particular consideration including the mean coefficient of determination (CDmean) and the mean prediction error variance (PEVmean) initially presented for contrasts between individuals (Laloë, 1993; Rincent et al., 2012), or clustering methods such as stratified sampling (Isidro et al., 2015; Akdemir and Isidro-Sánchez, 2019; Guo et al., 2019) and partitioning around medoids (PAM) (Guo et al., 2019). The CDmean, PEVmean and PAM criteria are now routinely used for TRS optimization, especially when the TRS size is small (Akdemir and Isidro-Sánchez, 2019). As part of the optimization process, the population structure plays a key role as it impacts

both the performance of the optimization methods (Isidro et al., 2015) and the GP predictive ability.

In a population stratified into genetics groups, when the same genetic groups are found within the TRS and the TS, the differences in means between groups are often implicitly taken into account by the model and contribute to the predictive ability (Guo et al., 2014; Rio et al., 2019). Conversely, when targeting a group-specific TS, training a model on a different group can dramatically limit the predictive ability, as shown in dairy and beef cattle (Olson et al., 2012; Chen et al., 2013) or maize (Technow et al., 2013; Lehermeier et al., 2014). As proposed by de Roos et al. (2009), genetic groups can also be combined into generic muti-group TRSs that show a good predictive ability regardless of the target population, as shown in dairy cattle (Brøndum et al., 2011; Pryce et al., 2011; Zhou et al., 2013), maize (Technow et al., 2013; Rio et al., 2019) or soybean (Duhnen et al., 2017). Several models have been proposed that explicitly account for genetic structure such as modeling genetic covariances between individuals from different groups by adapting multi-trait models (Karoui et al., 2012; Lehermeier et al., 2015).

The use of genomics in oat breeding is rather limited compared to other cereals like maize, wheat or rice, due to the scarcity of available tools. It can be explained by the complexity of its allo-hexaploid genome (2n=6x=42) with high content in repetitive sequences (Yan et al., 2016). Nevertheless, thanks to the efforts of the oat community over the past few years, several genome tools have been developed such as the Illumina 6K gene chip (Tinker et al., 2014), genotyping-by-sequencing (GBS) (Huang et al., 2014; Bekele et al., 2018), and a consensus map (Chaffin et al., 2016). Those tools enabled many genetic studies and breeding applications (Esvelt Klos et al., 2016; Tumino et al., 2016; Yan et al., 2016; Bjørnstad et al., 2017; Tumino et al., 2017; Carlson et al., 2019; Kebede et al., 2019; Sunstrum et al., 2019; Isidro-Sánchez et al., 2020a,b; Yan et al., 2020). More recently, the draft of the hexaploid *Avena sativa* genome sequence: OT3098 v1 - PepsiCo[1] and the sequence of two diploid oat genomes: *Avena Atlantica* and *Avena Eriantha* (Maughan et al., 2019) have been released. They will open a new frontier for the study of the oat genome and for the development of genomics-assisted breeding.

A few studies have focused on the application of GP and genomic selection (GS) in oat (Asoro et al., 2011, 2013; Bekele et al., 2018; Mellers et al., 2020; Haikka et al., 2020a,b). In these studies, the objectives included the comparison of GS to traditional phenotypic and marker-assisted selection for β-glucans (Asoro et al., 2013), GP of heading date using SNPs and tag-levels haplotype markers (Bekele et al., 2018), GP of agronomic traits and Fusarium head blight in an oat commercial breeding program (Haikka et al., 2020b,a), and the implementation of GP within a winter oat biparental cross (Mellers et al., 2020). These empirical GS applications have demonstrated the effective use of GS within breeding populations to accelerate oat breeding. Nevertheless, there is a lack of experimental studies focusing on the efficiency of GP and TRS optimization in highly structured oat populations.

In this paper, a structured Mediterranean oat population, including both white and red oat inbred lines, was evaluated for agronomic traits. The objectives were to (i) evaluate the predictive ability of different GP models, (ii) optimize TRSs

[1] https://wheat.pw.usda.gov/GG3/graingenes_downloads/oat-ot3098-pepsico

using different methods, and (iii) evaluate the impact of genetic structure on both the GP predictive ability and the performance of optimization methods.

Materials and Methods

Genetic material and genotypic data

Genetic material consists of a collection of 709 cultivated oat (*Avena sativa*) inbred lines, including landraces from the Mediterranean area, provided by the "Centro de Recursos Filogenéticos" of the "Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria" (INIA, Madrid, Spain) and the United States Department of Agriculture (Washington, USA), along with cultivars and breeding lines provided by different institutions, as presented in Sánchez-Martín et al. (2014) and Canales et al. (2021a). All individuals were genotyped using GBS-SNP markers as detailed in Canales et al. (2021a). Genotypes were pooled into libraries of 96 genotypes at the genomic platform of the McGill University (Canada), following the PstI-MspI method (Huang et al., 2014). Each GBS library was sequenced on a single line of a HiSeq 2500 at the "Plateforme d'Analyses Génomiques of the Institut de Biologie Intégrative et des Systèmes" of the "Université Laval" (Quebec City, Canada). Raw FASTQ sequences were processed using the Haplotag pipeline (Tinker et al., 2016). After a filtering on the minor allele frequency ($> 5\%$), the heterozygosity rate ($< 20\%$) and the percentage of missing values ($< 50\%$), a total of 17,288 bi-allelic SNP markers corresponding to 12,418 tags were obtained. Ten individuals were also discarded due to a large heterozygosity rate and/or a large percentage of missing values, leaving 609 individuals for subsequent analyses. Missing values at SNPs were then imputed using the multivariate normal expectation maximization algorithm (Poland and Rife, 2012) implemented in the R package rrBLUP (Endelman, 2011). The marker dataset is available at Dryad Data (Canales et al., 2021b) and sequence read data are available from NCBI SRA archive as BioProject ID PRJNA693576 (http://www.ncbi.nlm.nih.gov/bioproject/693576).

Structure analysis

A structure analysis was performed using the STRUCTURE software (Pritchard et al., 2000) for a number of genetic groups Q ranging from 2 to 5 and using the admixture model with correlated allele frequencies between groups (Falush et al., 2003). Each analysis consisted of 10,000 MCMC iterations and a burn-in of 1,000 iterations. Admixture barplots are presented in Fig. 1A for $Q = 3$ and in Online Resource Fig. S1 for other values of Q. The population could be separated into two groups corresponding to the two oat subspecies forming the population (*sativa* and *byzantina*), further referred to as Byzantina and Sativa. As the Sativa group was mainly structured into two sub-groups, we considered three genetic groups for further analyses by assigning individuals using their maximal admixture coefficient: Byzantina (257 lines), Sativa_A (243 lines) and Sativa_B (199 lines). Global position system coordinates were available for most individual accessions (Canales et al., 2021a) and revealed a relationship between the site where individuals were collected and the genetic group to which they were assigned (Fig. 1B). The Evanno method

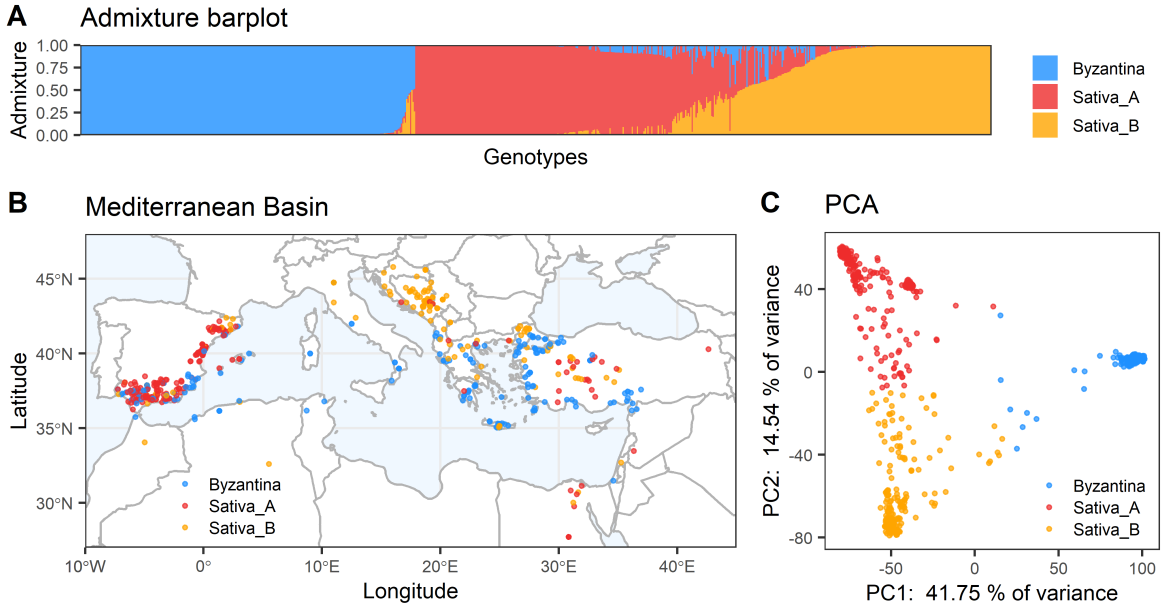


Fig. 1 Graphs illustrating the structure analysis performed on the oat dataset using STRUCTURE and showing the existence of three genetic groups: Byzantina (257 individuals), Sativa_A (243 individuals) and Sativa_B (199 individuals). Graphs include (**A**) an admixture barplot showing the admixture proportions of each individual, (**B**) a map of the Mediterranean Basin with dots indicating the location where individuals have been collected (Canales et al., 2021a), and (**C**) the two principal components of a principal component analysis (PCA) performed on genomic data. For graphs **B** and **C**, dots were colored according to the group to which each individual was assigned based on its maximal admixture coefficient.

was applied was applied and supported the use of $Q = 2$ groups due to the high genetic divergence between Byzantina and Sativa subspecies (Online Resource Figure S2). Like in Canales et al. (2021a), where a similar structure analysis was performed on this same dataset with very consistent results, and a previous study based on SSR markers (Montilla-Bascón et al., 2013), we subdivided the Sativa group into two groups, which was supported by a substantial proportion of variance explained by the axis differentiating these groups (14.54%) for a principal component analysis (PCA) performed on the genomic data (Fig 1C). Note that, unlike in Canales et al. (2021a), we did not define additional categories for admixed individuals.

Phenotypic analysis

The oat collection was evaluated in 2017 and 2018 in Cordoba (Spain) with an altitude of 90 m and light clay calcic cambisol soil, and in 2018 in Santaella (Spain) with an altitude of 238 m, and a light clay eutric gleysol soil, forming three distinct trials (Co17, Co18, and Sa18). Each trial was an alpha lattice square design with three replicates using the cultivar Patones as check. Four agronomic traits were evaluated: heading time (Heading) in growing degree-days, plant height (Height)

in cm, vegetative biomass (Biomass) in t/ha and grain yield (Yield) in t/ha, see Canales (2019) and Canales et al. (2021a) for more details. The contribution of genotype-by-environment (GxE) interactions to the phenotypic variance and the broad-sense heritability were investigated using the following model:

$$Y_{ijkr} = \mu + \alpha_k + \beta_j + G_{ik} + (G \times \beta)_{ijk} + \gamma_{rj} + E_{ijkr},$$

where Y_{ijkr} is the phenotype of individual i from group k in block r of trial j, μ is the global intercept, α_k is the effect of group k with $k \in \{By, SaA, SaB\}$ (By: Byzantina, SaA: Sativa_A and SaB: Sativa_B), β_j is the effect of trial j, G_{ik} is the random genotypic effect of individual i from group k with $G_{ik} \sim \mathcal{N}\left(0, \sigma^2_{G_k}\right)$ independent, $(G \times \beta)_{ijk}$ is the random genotype-by-environment (GxE) effect of individual i from group k in trial j with $(G \times \beta)_{ijk} \sim \mathcal{N}\left(0, \sigma^2_{G\times\beta)_k}\right)$ independent, γ_{rj} is the effect of block r in trial j, and E_{ijkr} is the error of individual i from group k in block r of trial j with $E_{iijkr)} \sim \mathcal{N}\left(0, \sigma^2_E\right)$ independent and identically distributed. All random effects are assumed to be independent of each other. Model parameters were estimated using the R package "MM4LMM" (Laporte and Mary-Huard, 2020). The group-specific means were calculated as following:

$$\mu_k = \mu + \alpha_k + \frac{1}{J}\sum_{j=1}^{J}\beta_j + \frac{1}{JR}\sum_{j=1}^{J}\sum_{r=1}^{R}\gamma_{rj},$$

where $J = 3$ is the number of trials and $R = 3$ is the number of blocks in each trial. The group-specific broad-sense heritabilities were calculated as following:

$$H^2_k = \frac{\sigma^2_{G_k}}{\sigma^2_{G_k} + \frac{1}{J}\sigma^2_{G\times\beta_k} + \frac{1}{JR}\sigma^2_E},$$

Least-square means (LS-means) of each individual (Y^*_{ik}) were computed based on the same model with G_{ik} and $(G \times \beta)_{ijk}$ as fixed effects using:

$$Y^*_{ik} = \mu_k + G_{ik} + \frac{1}{J}\sum_{j=1}^{J}(G \times \beta)_{ijk},$$

and were further referred to as phenotypes for GP analyses.

Genomic prediction models

In this study, the first four GP models (GBLUP, MGBLUP, Bayes-B and RKHS) can be written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is the vector of reference phenotypes (i.e., the LSmeans), $\boldsymbol{X}$ is the design matrix for fixed effects, $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{Z}$ is the incidence matrix linking phenotypes to breeding values, $\boldsymbol{g}$ is the vector of breeding values and $\boldsymbol{e}$ is the vector of errors. All models assume independence between $\boldsymbol{g}$ and $\boldsymbol{e}$.

GBLUP A standard additive GBLUP model was applied using the R package "rrBLUP" (Endelman, 2011) with the following assumptions: $\boldsymbol{\beta} = \mu$ is the overall mean, $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{K}\sigma^2_G)$, $\boldsymbol{K}$ is the kinship matrix, σ^2_G is the genetic variance, $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}\sigma^2_E)$, $\boldsymbol{I}$ is the identity matrix and σ^2_E is the error variance. The kinship between individuals i and j ($K_{i,j}$) was computed following VanRaden (2008):

$$K_{i,j} = \frac{\sum_{m=1}^{M}(W_{im} - f_m)(W_{jm} - f_m)}{\sum_{m=1}^{M} f_m(1 - f_m)}, \tag{2}$$

where M is the number of markers, W_{im} is the genotypic score of individual i at locus m (coded 0; 0.5; 1) and f_m is the frequency of allele "1" at locus m estimated on the whole dataset.

MGBLUP A multi-group GBLUP (MGBLUP) model (Lehermeier et al., 2015) was applied using the R package "MTM"[2]. This model consists of adapting a multi-trait model to the analysis of one trait in different groups, which allows for group-specific genetic variances and specific genetic covariances between groups. In this model, $\boldsymbol{\beta} = (\mu_{By}, \mu_{SaA}, \mu_{SaB})^T$ is the vector of group-specific means, $\boldsymbol{g} = \begin{bmatrix} \boldsymbol{g}^*_{By} \\ \boldsymbol{g}^*_{SaA} \\ \boldsymbol{g}^*_{SaB} \end{bmatrix}$ is the expanded vector of breeding values of each individual in each group of size of $3N$, N being the number of individuals, where: $\begin{bmatrix} \boldsymbol{g}^*_{By} \\ \boldsymbol{g}^*_{SaA} \\ \boldsymbol{g}^*_{SaB} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2_{G_{By}} & \sigma_{G_{By,B}} & \sigma_{G_{By,SaB}} \\ \sigma_{G_{By,SaA}} & \sigma^2_{G_{SaA}} & \sigma_{G_{SaA,SaB}} \\ \sigma_{G_{By,SaB}} & \sigma_{G_{SaA,SaB}} & \sigma^2_{G_{SaB}} \end{bmatrix} \bigotimes \boldsymbol{K}\right)$, with $\sigma_{G_{X,Y}}$ being the genetic covariance between groups X and Y (the letters X, Y are further used as group names when not specifically designating given groups), $\boldsymbol{K}$ the same kinship matrix computed following Eq. (2), and $\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_{By} \\ \boldsymbol{e}_{SaA} \\ \boldsymbol{e}_{SaB} \end{bmatrix}$ is the vector of errors of size N where: $\begin{bmatrix} \boldsymbol{e}_A \\ \boldsymbol{e}_B \\ \boldsymbol{e}_C \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{I}_{By}\sigma^2_{E_{By}} & 0 & \\ 0 & \boldsymbol{I}_{SaA}\sigma^2_{E_{SaA}} & 0 \\ 0 & 0 & \boldsymbol{I}_{SaB}\sigma^2_{E_{SaB}} \end{bmatrix}\right)$, $\boldsymbol{I}_X$ is the identity matrix with dimensions equal to the number of observations from group X and $\sigma^2_{E_X}$ is the error variance in group X. The choice of hyper-parameters for the inference was done following Lehermeier et al. (2015).

Bayes-B The Bayesian shrinkage regression Bayes-B proposed by Meuwissen et al. (2001) was applied using the R package "BGLR" (Pérez and de los Campos, 2014). In Bayes-B, only a proportion of a markers can have a non-zero effect with a variance specific to each marker. This modeling represents genetic architectures for which some SNPs are not associated to any QTL while others are associated to QTL with potentially large effects. In this model, $\boldsymbol{g} = \boldsymbol{W}\boldsymbol{u}$, $\boldsymbol{W}$ is the centered genotyping matrix and $\boldsymbol{u}$ is the vector of marker effects where the prior distribution of each u_m is the following mixture distribution:

$$\mathrm{P}\left(u_m|\pi\right) = \begin{cases} 0 & \text{with probability } \pi, \\ t(0, \nu, S^2) & \text{otherwise,} \end{cases}$$

[2] available at https://github.com/QuantGen/MTM

where π is the proportion of marker with null effect, $t(0, \nu, S^2)$ is the scaled-t distribution with ν and S^2 being the number of degrees of freedom and the scale parameter, respectively. Other terms are identical to those of GBLUP.

RKHS The reproducing kernel Hilbert space (RKHS) semiparametric approach for genomic prediction (Gianola et al., 2006; Gianola and van Kaam, 2008) was applied using R package "BGLR" (Pérez and de los Campos, 2014). This approach combines a classical additive genetic model with a kernel function which converts predictor variables into set of distances among observations. The RKHS model based on a Gaussian kernel has been demonstrated to capture epistatic effects between markers (Jiang and Reif, 2015). In this model, $\boldsymbol{g} = \boldsymbol{K}_h\boldsymbol{\alpha}$, $\boldsymbol{K}_h$ is the matrix of kernel entries and $\boldsymbol{\alpha}$ is the vector of individual effects. Other terms are identical to those of GBLUP. The kernel function implemented here was a Gaussian kernel:

$$\boldsymbol{K}_h(\boldsymbol{W}_i, \boldsymbol{W}_j) = e^{-hd_{i,j}},$$

where h is a smoothing parameter and $d_{i,j}$ is the marker-based Euclidean distance between individuals i and j. The value of the smoothing parameter was chosen following the kernel averaging method proposed in Pérez and de los Campos (2014).

Random Forest The tree-based machine learning approach called Random Forest (Breiman, 2001; Chen and Ishwaran, 2012) was applied using the R package "randomForest" (Liaw and Wiener, 2007). The grouping property of trees enables the Random Forest to adequately deal with correlations and interactions between predictor variables (Chen and Ishwaran, 2012). In this approach, the vector $\boldsymbol{y}$ of phenotypes was used as the vector of response variable while the centered genotyping matrix $\boldsymbol{W}$ was used as the matrix of predictor variables.

For the GP models based on a Bayesian inference (MGBLUP, Bayes-B and RKHS), 10,000 MCMC iterations were considered with a burn-in of 1,000 iterations and a thinning of 3 (i.e., one out of three samples were conserved to compute posterior means).

Evaluation of the predictive ability of genomic prediction

The precision of the models was evaluated using three different cross-validation (CV) procedures where the predictive ability was calculated by correlating the predictions of breeding values of the TS to the reference phenotypes (i.e., the LSmeans).

The first CV procedure, referred to as holdout cross-validation (HO-CV), was performed by repeatedly splitting (x 100) the oat population into a TRS and a TS with proportions being 4/5 and 1/5, respectively. This CV procedure makes it possible to study the level of precision that can be obtained when neglecting the role of genetic structure.

The second CV procedure, referred to as leave-one-out cross-validation (LOO-CV), was performed by predicting the breeding value of each individual using a model trained on all the remaining individuals. It allowed a joint graphic representation of the quality of prediction of all individuals depending on their genetic group.

The third CV procedure, referred to as structured-holdout cross validation (SHO-CV), allowed to study the impact of genetic structure on the predictive ability, as presented in Rio et al. (2019). In SHO-CV, group-specific TSs of 49 individuals were predicted using a model calibrated on 150 other individuals. Depending on the scenario, the training set included either members of a single group (e.g., 150 individuals from the Byzantina group), or of the three groups in balanced proportions (i.e., 50 individuals from each group). The sampling was repeated 100 times for each scenario.

Training set optimization

For TRS optimization, the oat population was repeatedly split (x 30) into a candidate set (CS) of N_{CS} individuals and a TS of N_{TS} individuals (Fig. 2). Both the CS and the TS were defined differently depending on the optimization scenario (see below). The TRS individuals were then selected among CS individuals based on PEVmean, CDmean or PAM (see below) and using a genetic algorithm implemented in the R package "STPGA" (Akdemir, 2017), using 500 iterations. For each TS, 30 random TRS were sampled as a benchmark. For each trait, the breeding values of TS individuals were then predicted using a GBLUP model trained on TRS individuals, and the predictive ability was calculated to compare optimization methods. Three different criteria were considered:

PEVmean and CDmean The PEVmean and CDmean optimization criteria are derived from the GBLUP linear mixed model. The PEVmean criterion is defined as the mean of the predictor error variance over a set of individuals, where each PEV_i can be computed as:

$$PEV_i = \mathrm{Var}(\boldsymbol{g}_i - \widehat{\boldsymbol{g}}_i) = \left(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^T + \boldsymbol{K}^{-1}\lambda\right)^{-1}_{i,i} \times \sigma^2_E,$$

where $\boldsymbol{g}_i$ is the breeding value of i, $\widehat{\boldsymbol{g}}_i$ is the best linear unbiased prediction (BLUP) of $\boldsymbol{g}_i$, $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is an orthogonal projector on the subspace spanned by the columns of $\boldsymbol{X}$ where $(\boldsymbol{X}^T\boldsymbol{X})^-$ is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$ (Laloë, 1993), and $\lambda = \dfrac{\sigma^2_E}{\sigma^2_G}$. All other terms correspond to those described in the GBLUP model. The CDmean criterion is defined as the mean of the coefficient of determination (i.e., the square correlation between the breeding value of an individual and its corresponding prediction) where each individual CD can be computed as:

$$CD_i = \mathrm{cor}(\boldsymbol{g}_i, \widehat{\boldsymbol{g}}_i)^2 = \frac{\left(\boldsymbol{K} - \lambda\left(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^T + \boldsymbol{K}^{-1}\lambda\right)^{-1}\right)_{i,i}}{\boldsymbol{K}_{i,i}}$$

In this study, both PEVmean and CDmean optimizations were "targeted" (Akdemir and Isidro-Sánchez, 2019), meaning that the criteria were computed directly over the TS individuals. Note that finding the best TRS is be done by minimizing PEVmean while it is done by maximizing CDmean.

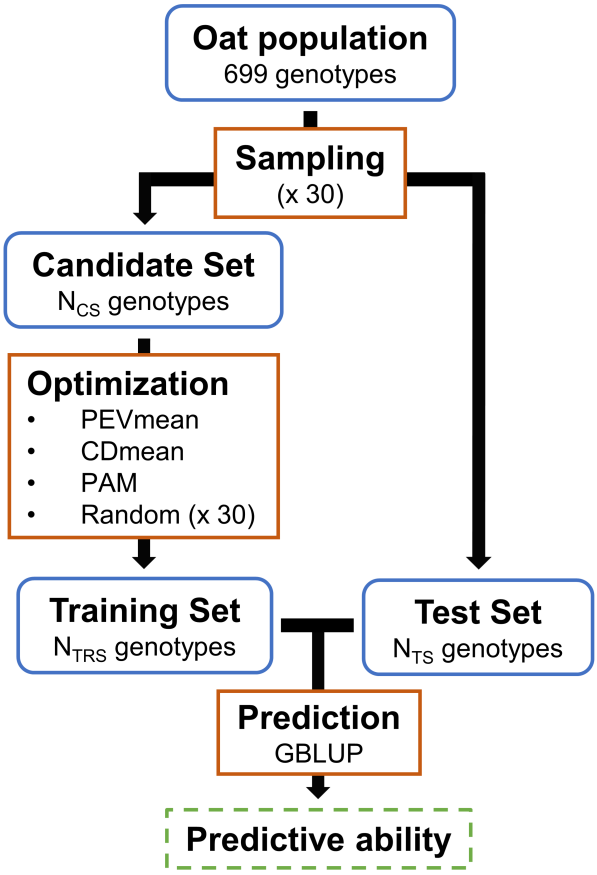


Fig. 2 Diagram illustrating the training set (TRS) optimization procedure where the oat population was repeatedly split (x 30) into a candidate set (CS) of N_{CS} individuals and a test set (TS) of N_{TS} individuals. Training set individuals were selected among CS individuals using a genetic algorithm (STPGA) and different methods: (i) the mean of the prediction error variance (PEVmean) of the TS, (ii) the mean of the coefficient of determination (CDmean) of the TS and (iii) partitioning around medoids (PAM). For each TS, 30 random TRS were sampled as a benchmark. The breeding values of TS individuals were then predicted using a GBLUP model trained on TRS individuals for each trait, before calculating the predictive ability for validation.

PAM Partitioning around medoids is a clustering method that classifies individuals into clusters by minimizing the sum of dissimilarities between the individuals of each cluster, and designating a central individual, or medoid of that cluster (Kaufman and Rousseeuw, 1987). The application of PAM to TRS optimization and genomic data was first presented by Guo et al. (2019).

Two optimization scenarios were considered in this study:

Standard optimization Test sets of 99 individuals were randomly sampled (i.e., by neglecting genetic structure). The CS consisted of all the remaining individuals. Optimized TRSs were selected for a gradient of TRS sizes (i.e., $N_{TRS} \in \{10, 30, 60, 100, 200, 300, 600\}$). Note that when $N_{TRS} = 600$, all CS individuals are included in the TRS.

Structure-based optimization. We sampled group-specific TSs of 49 individuals (e.g., 49 individuals from the Byzantina group). The CS consisted of 150 remaining individuals from each genetic group. Optimized TRSs were selected for a gradient of TRS sizes (i.e., $N_{TRS} \in \{9, 15, 30, 60, 90, 120, 150\}$). Note that $N_{TRS} = 150$ corresponds to the limit for which only individuals from the same groups as TS individuals can be selected during the optimization procedure.

Results

Phenotypic characterization of the population

Our oat population has been evaluated for four agronomic traits: Heading, Height, Biomass and Yield. As expected, agronomic performance of the collection varied overall in the different environments. In addition to the different altitudes and soil structure of the different sites already stated, Co17 was characterised by a mean maximum T^a of 21.49°C, a mean minimum of 7.51°C and a rainfall of 415 mm during the growing season. Co18 was slighty warmer with a mean maximum T^a of 21.50°C, a mean minimum of 8.46°C and a rain of 497 mm during the growing season. Sa18, was slighty colder and rainier than any of the environments of Cordoba, with a mean maximum T^a of 20.07 °C, a mean minimum of 7.64°C and a rainfall of 513 mm during the growing season (Online Resource Table S1). In these environments, mean Heading values ranged from 174 days to heading at Co18 to 150 days at Co17, with a minimum of 118 days and a maximum of 200 days to heading. Mean Height values ranged from 115 cm at Co18 to 139 cm at Sa18 with the shortest accession reaching 68 cm and the longest reaching 191 cm. Regarding Biomass, mean values ranged from 6832 kg/ha at Co17 to 2893 at Co18 with minimum values of 190 kg/ha and maximum of 11766 Kg/ha. Mean yield ranged from 2326 Kg/ha at Sa18 to 983 kg/ha at Co18, with the highest yielding accessions reaching 5326 Kg/ha and the lowest one yielding 180 Kg /ha. All evaluated traits showed moderate to high broad-sense heritabilities, with variations depending on the genetic group (Table 1). For instance, a broad-sense heritability of 0.57, 0.74 and 0.55 were estimated for Biomass in the Byzantina, Sativa A and Sativa B groups, respectively. Those variations could be connected to the differences in genetic variances observed between groups (i.e., the larger the group-specific genetic variance $\sigma^2_{G_X}$, the higher the group-specific broad-sense heritability H^2_X). The GxE variances were not large for most traits and also variable depending on the genetic group. In terms of means, the performances of each genetic group were comparable for most traits, but the Byzantina group outperformed both Sativa groups for Yield and Biomass.

Table 1 Group-specific means, variances and broad-sense heritabilities estimated in the phenotypic analysis for Heading (in 100 growing degree-days (GDD)), Height (in cm), Biomass (in t/ha) and Yield (in t/ha).

	Heading	**Height**	**Biomass**	**Yield**
μ_{By}	9.62	131.33	5.91	1.85
μ_{SaA}	10.71	124.63	5.19	1.43
μ_{SaB}	10.60	133.54	5.25	1.25
$\sigma^2_{G_{By}}$	0.76	76.43	0.95	0.14
$\sigma^2_{G_{SaA}}$	2.01	113.13	2.38	0.46
$\sigma^2_{G_{SaB}}$	1.04	290.85	1.38	0.14
$\sigma^2_{(G\times\beta)_{By}}$	0.04	0.00	0.52	0.11
$\sigma^2_{(G\times\beta)_{SaA}}$	0.28	79.98	0.84	0.12
$\sigma^2_{(G\times\beta)_{SaB}}$	0.18	11.44	1.72	0.06
σ^2_E	0.27	189.33	4.95	0.32
H^2_{By}	0.95	0.78	0.57	0.65
H^2_{SaA}	0.94	0.70	0.74	0.86
H^2_{SaB}	0.92	0.92	0.55	0.72

Predictive ability of genomic prediction models

The GP predictive ability was evaluated using HO-CV for different GP models: GBLUP, Bayes-B, RKHS, Random-Forest and MGBLUP (Fig. 3). The mean predictive abilities averaged over all models were 0.87, 0.72, 0.63 and 0.81 for Heading, Height, Biomass and Yield, respectively. In terms of model comparison, GBLUP, RKHS and MGBLUP showed similar predictive abilities for all traits. Bayes-B showed similar predictive ability to GBLUP (our reference model) for Heading, Height and Yield, and higher predictions than GBLUP for Biomass (0.68 vs. 0.60).

Training set optimization targeting random test sets

Three TRS optimization methods (PEVmean, CDmean and PAM) were compared to random sampling for a gradient of TRS sizes and validated using the four traits (Fig. 4). For all traits, the predictive capability increased with the size of the TRS, but at a rate that decreased as the size of the TRS increased. Optimizations of TRS based on PEVmean and CDmean performed very similarly and generally allowed for higher gains compared to the optimization based on PAM. For Height and $N_{TRS} = 100$, TRSs selected by CDmean and PEVmean showed a mean predictive ability of 0.70 compared to 0.61 and 0.60 for PAM and random sampling, respectively. The gains obtained with the optimization based on PAM were very variable depending on the trait, and even led to a substantial loss in predictive ability for Biomass (e.g., for $N_{TRS} = 100$, TRSs selected by PAM showed a mean predictive ability of 0.35 compared to 0.43 for random sampling). The mean group proportions within selected TRSs did not reveal major differences between optimization methods (Online Resource Fig. S3).

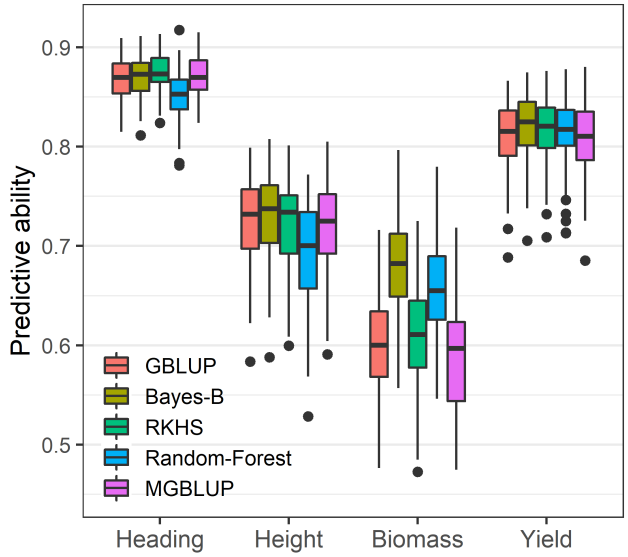


Fig. 3 Boxplots of predictive abilities obtained for each trait using holdout cross-validations (x 100 replicates) and five GP models: GBLUP, Bayes-B, RKHS, Random-Forest and multi-group GBLUP (MGBLUP).

Genetic structure and predictive ability

The impact of genetic structure on the predictive ability was first investigated graphically using LOO-CV (Online Resource Fig. S4). This approach confirmed the ability of the GBLUP model to predict differences in mean between groups and suggested an ability of the model to predict beyond this simple effect of genetic structure. We then investigated the impact of genetic structure using SHO-CV to explore within/across/multi-group scenarios (Fig. 5). In general, to predict a group-specific TS, the best strategy was to train a model using individuals from the same genetic group. The worst predictive abilities were obtained by training a model using the Byzantina group to predict any of the Sativa groups, or vice versa. Interestingly, a negative mean correlation of -0.55 was obtained when predicting the Sativa group using the Byzantina group for Heading. In general, the predictive abilities obtained with across-group scenarios between the Sativa_A and Sativa_B groups were moderate but not always symetric. For instance, with Yield, a predictive ability of 0.58 was achieved when training a model on the Sativa_B group to predict the Sativa_A group, while a predictive ability of 0.22 was obtained for the opposite scenario. Multi-group TRSs always allowed for moderate to high predictive abilities regardless of the targeted TS.

Training set optimization targeting group-specific test sets

The impact of genetic structure on TRS optimization was investigated using structure-based optimization scenarios (Fig. 6). Two optimization methods (CD-

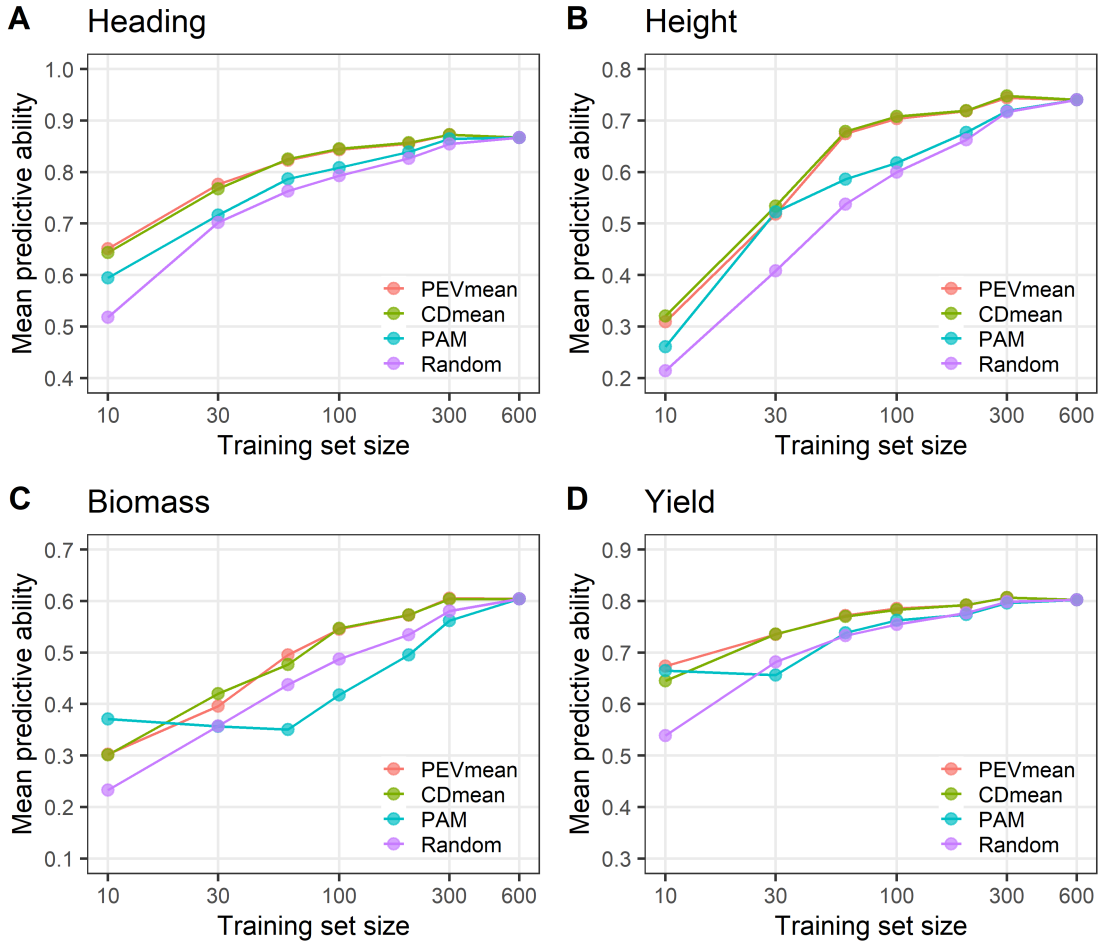


Fig. 4 Plots of mean GBLUP predictive abilities over random test sets (TSs) of 99 individuals (x 30 replicates) according to the size of the training set (TRS) for (**A**) Heading, (**B**) Height, (**C**) Biomass and (**D**) Yield. Different optimization methods were compared: PEVmean, CDmean and PAM, along with random sampling as a benchmark (x 30 replicates for each TS).

mean and PAM) were compared to random sampling for selecting TRSs that best predict group-specific TSs. The optimization based on CDmean always led to higher gains compared to PAM. The differences between optimization methods could largely be explained by the ability of CDmean to preferentially select individuals from the same genetic group as the one in the TS, unlike PAM or random sampling (Online Resource Fig. S5). However, the observed gains were very variable depending on the group-specific TS and the trait. For $N_{TRS} = 30$ and Yield, the gains in predictive ability from selecting TRSs using CDmean compared to random was +0.35 for Byzantina TSs compared to +0.03 for Sativa_A TSs (Fig. 6).

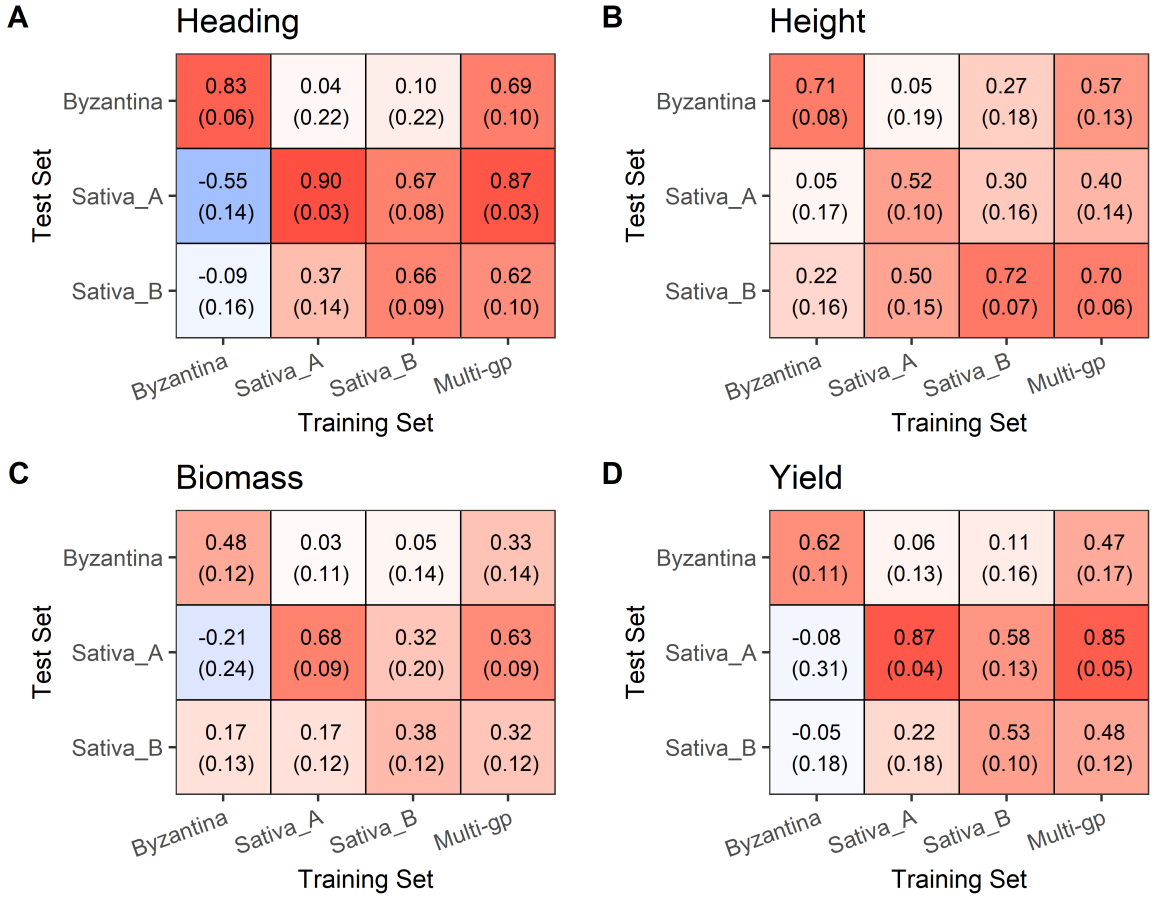


Fig. 5 Heatmaps of mean GBLUP predictive abilities obtained using the structured holdout cross-validations (x 100 replicates) for (**A**) Heading, (**B**) Height, (**C**) Biomass and (**D**) Yield. Group-specific training sets (TRSs) and multi-group TRSs of 150 individuals are indicated on the x axis while group-specific test sets of 49 individuals are indicated on the y axis. Standard deviations are shown between brackets.

Discussion

Oat genomics-assisted breeding in the Mediterranean basin

The GP predictive ability of oat agronomic and phenological traits has only been the subject of a few studies in the last past years (Asoro et al., 2011; Bekele et al., 2018; Haikka et al., 2020b,a; Mellers et al., 2020). The moderate to high predictive abilities obtained in our study are comparable to the ones obtained in the latter studies. It confirms the value of GBS-SNP markers as a genotyping technology for implementing GS in oat breeding programs, as proposed by Huang et al. (2014). The similar performances achieved by the different GP models tested in this study is a common feature in the GP literature (Heslot et al., 2012; de los Campos et al., 2013). The large gain in predictive ability obtained by applying Bayes-B for

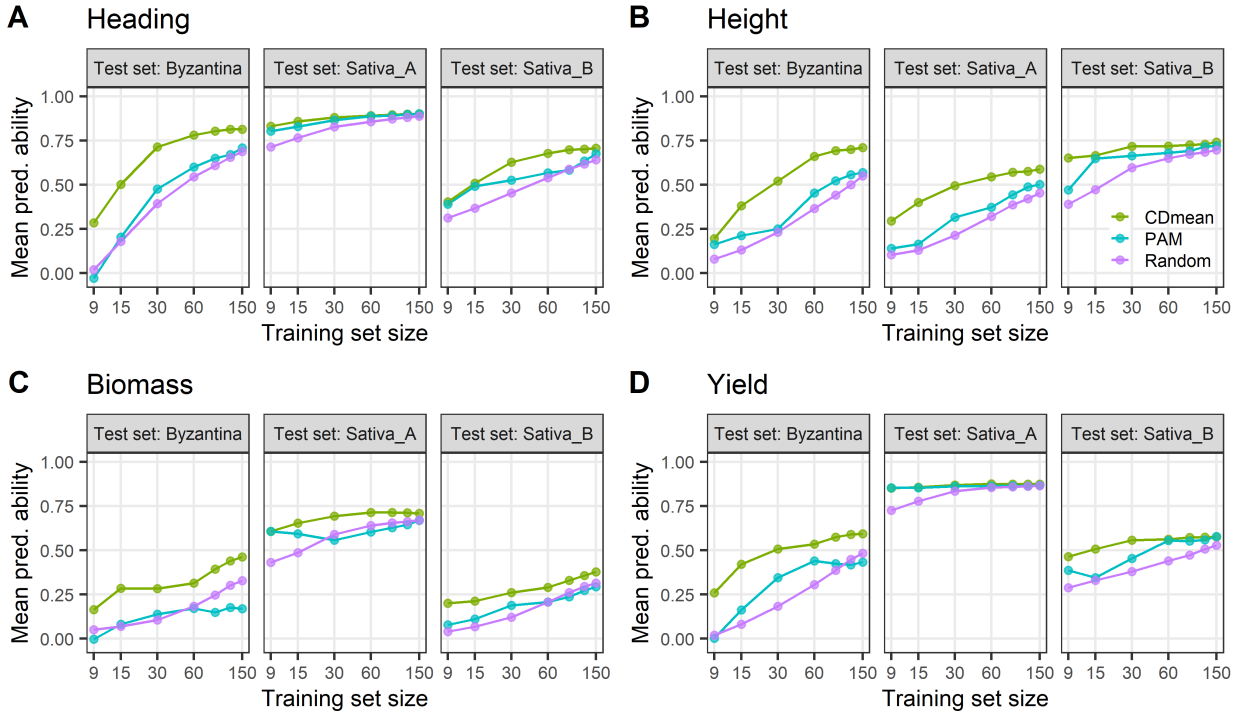


Fig. 6 Plots of mean GBLUP predictive abilities (Mean pred. ability) over group-specific test sets (TS) of 49 individuals (x 30 replicates) according to the size of the training set (TRS) for (**A**) Heading, (**B**) Height, (**C**) Biomass and (**D**) Yield. Two optimization methods were compared: CDmean and PAM, along with random sampling with equal group contributions in the TRS as a benchmark (x 30 replicates for each TS). Each candidate set consisted of 450 individuals, including 150 from each of the three genetic groups.

Biomass may result from the existence of a QTL with large effect segregating in the population that can better be accounted for using Bayes-B (Meuwissen et al., 2001; Pérez and de los Campos, 2014).

To improve the productivity of oats in the hot and dry environment of the Mediterranean basin, breeding programs must not only be based on the use of efficient tools like GP, but also on the introgression of favorable alleles from various sources of genetic diversity. Landraces harbour a great genetic potential for oat improvement as they are endued with a higher genetic variability compared to high-yielding cultivars (Montilla-Bascón et al., 2013; Sánchez-Martín et al., 2016; Winkler et al., 2016). The population evaluated in our study includes landraces and cultivars of both white and red oats. A restricted set of this broad germplasm has already been characterized in field trials in different countries of the Mediterranean rim (Sánchez-Martín et al., 2014) and showed potential to detect QTL for powdery mildew and crown rust resistance (Montilla-Bascón et al., 2015), as well as agronomic traits (Rispail et al., 2018). In our study, all genetic groups showed a substantial genetic variance and comparable means for all traits. We can reasonably assume that different QTL are involved in the trait genetic variability depending on the genetic group. Since population structure results from differences in allele frequencies

between groups, a QTL may indeed be polymorphic in one group and contribute to its genetic variance, while being fixed in another group. This suggests a great potential of both white and red oat landraces for harnessing new alleles to be introgressed into elite germplasm using pre-breeding methods based on GP (Gorjanc et al., 2016; Allier et al., 2020).

Genomic prediction in a highly structured population

When assessing the predictive ability of GP within structured populations, the observed predictive ability may result mostly from the ability of the model to predict differences in mean between groups (Guo et al., 2014; Rio et al., 2019). For all traits, the SHO-CV showed that the predictive ability was also moderate to high when considering group-specific TSs, meaning that it would be possible to identify the best individuals within each genetic group (Fig. 5). Our results were in concordance with previous studies on GP in structured populations (Rio et al., 2019). Thus, (i) a given group-specific TS was generally best predicted using a TRS including only individuals from the same genetic group, (ii) applying across-group predictions could highly depreciate the GP predictive ability, and (iii) a multi-group TRS showed a high predictive ability, regardless of the targeted TS.

The genetic dissimilarity between the two oat subspecies *sativa* and *byzantina* was illustrated by the difficulty to predict one of them using a model trained on the other (Fig. 5). The substantial negative predictive ability obtained when predicting the Sativa_A group using a model trained on the Byzantina group for Heading is uncommon in GP. It may result from the existence of QTL segregating in the population with effects of opposite signs depending on the genetic group. Interestingly, previous results based on the same data (Canales et al., 2021a) identified a marker associated with heading date (avgbs_cluster_1918.1) located in chromosome 1D in the hexaploid reference genome. Comparison of the region around the marker between the *byzantina* and *sativa* preliminary genome assemblies have identified the GAS4-like gibberellin responsive gene, and a gene with homology to miR172 as potential candidates genes (Canales et al., 2021a). Conversely to what was observed between Byzantina and Sativa groups, both Sativa_A and Sativa_B were able to predict each other with moderate predictive abilities. This difference can be explained by a greater genetic similarity between the Sativa groups compared to that observed between the Sativa and Byzantina groups (Fig. 1C). These results are in concordance with the existence of moderate population structure in white oat, as observed in previous studies (Asoro et al., 2011; Newell et al., 2011, 2012; Huang et al., 2014; Tumino et al., 2016; Esvelt Klos et al., 2016; Winkler et al., 2016; Bjørnstad et al., 2017; Haikka et al., 2020b,a; Isidro-Sánchez et al., 2020a). They also illustrate the need for the evaluation of population structure when applying GP to a broad diversity (Isidro et al., 2015; Guo et al., 2019). Defining a TRS must be done by selecting individuals from the same genetic groups as those represented in the target population to maximize the predictive ability of GPs. If the target population is not clearly identified, the best strategy is to define a generic TRS that include all genetic groups (de Roos et al., 2009; Rio et al., 2019).

Training set optimization in a highly structured population

Training set optimization methods like CDmean, PEVmean and PAM can be used to select a subset of individuals to be evaluated when budget limitations limit the possibility to evaluate all possible individuals through extensive field trials (see Akdemir and Isidro-Sánchez (2019) and references herein). Unlike previous TRS optimization results in highly structured population presented by Isidro et al. (2015), in our results CDmean and PEVmean allowed for substantial gains compared to random sampling for all traits. This might be explained by several differences regarding the methodology: (i) the criteria were computed directly on the TS using the targeted optimization recommended by Akdemir and Isidro-Sánchez (2019) rather than on the remaining candidates, and (ii) the optimization algorithm was the genetic algorithm implemented in the "STPGA" R package (Akdemir, 2017) rather than an exchange algorithm. Note also that the CD and the PEV computed were those associated with the prediction of each breeding value and not with a contrast between a set of breeding values. A possible extension of the CDmean and PEVmean optimization criteria could be to compute the CD and PEV associated with contrasts between the breeding value of each TS individual and the mean of the TS, as recommended by Rincent et al. (2017).

The performance of the optimization based on PAM proposed by Guo et al. (2019) was highly variable and could even lead to predictive abilities that were worse than those obtained by random sampling. Our results illustrate the superiority of using criteria that are directly connected to the quantity of interest (e.g., CD corresponds to the model-based square correlation between the breeding value of an individual and its prediction) rather than using heuristic approaches like PAM. The aim of the optimization based on PAM is indeed to maximize the GP predictive ability by maximizing the genetic distances between TRS individuals. The differences between PAM and CDmean/PEVmean optimizations show that a targeted optimization based on CDmean/PEVmean not only aims at maximizing the distances between TRS individuals, but also implicitly accounts for other criteria such as the minimization of the genetic distances between TRS and TS individuals (Pszczola et al., 2012; Albrecht et al., 2011; Clark et al., 2012). The superiority of the targeted CDmean compared to PAM was also shown by its ability to select the TRS according to the nature of the target. For instance, if the target consists only of Byzantina individuals, then a straightforward strategy would be to preferentially select Byzantina individuals to form the TRS. While this strategy is implicitly applied using a targeted CDmean optimization, the PAM optimization does not account for any information of the TS, and may lead to selecting individuals that are poorly connected to the targeted population.

Conclusion

This manuscript presents results on the GP of key agronomic traits in a diverse populations of Mediterranean oat cultivars and landraces. The consistency between the structure of the training population and the population to be predicted was key to the predictive ability of genomic predictions. Regarding TRS optimization, the superiority of CDmean and PEVmean compared to PAM was illustrated by their ability to adapt the representation of each genetic group according to those

represented in the population to be predicted. Our findings are useful for future studies that aims to implement genomics-assisted breeding tools in presence of high population structure in oat and other species.

Funding

This work was supported by the Spanish Ministry of Science and Innovation [PID2019-104518RB-100], (AEI/FEDER, UE) and the regional government through the AGR-253 group and the European Regional and Social Development Funds. LGS is holder of a FPI fellowship from the Spanish Ministry of Economy and Competitiveness [BES-2017-080152]. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 818144, and also the Severo Ochoa Program for Centres of Excellence in R&D. JIS was supported by the Beatriz Galindo Program (BEAGAL18/00115) from the Ministerio de Educación y Formación Profesional of Spain and the Severo Ochoa Program for Centres of Excellence in R&D from the Agencia Estatal de Investigación of Spain, grant SEV-2016-0672 (2017-2021) to the CBGP.

Author contribution statement

SR, JIS and EP conceived the study. FJC and GMB participated in the genotyping and phenotyping. SR and LGS performed statistical analyses. SR, LGS, JIS and EP drafted the manuscript. All authors discussed the results and reviewed the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical standards

The authors declare that the experiments comply with the current laws of the countries in which the experiments were performed.

References

Akdemir, D. (2017). Stpga: Selection of training populations with a genetic algorithm. *bioRxiv*.

Akdemir, D. and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports*, 9(1):1446.

Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, 47(1):38.

566 Albrecht, T., Wimmer, V., Auinger, H.-J., Erbe, M., Knaak, C., Ouzunova, M.,
567 Simianer, H., and Schön, C.-C. (2011). Genome-based prediction of testcross
568 values in maize. *Theoretical and Applied Genetics*, 123(2):339.
569 Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., and Charcosset, A. (2020).
570 Optimized breeding strategies to harness genetic resources with different perfor-
571 mance levels. *BMC Genomics*, 21(1).
572 Alvarenga, A. B., Veroneze, R., Oliveira, H. R., Marques, D. B., Lopes, P. S.,
573 Silva, F. F., and Brito, L. F. (2020). Comparing alternative single-step gblup
574 approaches and training population designs for genomic evaluation of crossbred
575 animals. *Frontiers in genetics*, 11:263.
576 Asoro, F., Newell, M., Beavis, W., Scott, P., Tinker, N., and Jannink, J.-L. (2013).
577 Genomic, marker-assisted, and pedigree-blup selection methods for beta-glucan
578 concentration in elite oat. *Crop Sci.*, 53:1894–1906.
579 Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J.-L. (2011).
580 Accuracy and training population design for genomic selection on quantitative
581 traits in elite north american oats. *The Plant Genome*, 4(2):132–144.
582 Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J., and Tinker, N. A. (2018).
583 Haplotype-based genotyping-by-sequencing in oat genome research. *Plant*
584 *Biotechnology Journal*, 16(8):1452–1463.
585 Berro, I., Lado, B., Nalin, R. S., Quincke, M., and Gutiérrez, L. (2019). Training
586 population optimization for genomic selection. *The Plant Genome*, 12(3):1–14.
587 Bjørnstad, Å., He, X., Tekle, S., Klos, K., Huang, Y.-F., Tinker, N. A., Dong, Y.,
588 and Skinnes, H. (2017). Genetic variation and associations involving fusarium
589 head blight and deoxynivalenol accumulation in cultivated oat (avena sativa l.).
590 *Plant Breeding*, 136(5):620–636.
591 Brandariz, S. P. and Bernardo, R. (2018). Maintaining the accuracy of genomewide
592 predictions when selection has occurred in the training population. *Crop Science*,
593 58(3):1226–1231.
594 Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.
595 Brøndum, R., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse,
596 W., and Lund, M. (2011). Reliabilities of genomic prediction using combined
597 reference data of the nordic red dairy cattle populations. *Journal of Dairy*
598 *Science*.
599 Canales, F. J. (2019). *Improving oat for adaptation to Mediterranean environments*.
600 PhD thesis, Universidad de Córdoba.
601 Canales, F. J., Montilla-Bascón, G., Bekele, W. A., Howarth, C. J., Langdon,
602 T., Rispail, N., Tinker, N. A., and Prats, E. (2021a). Population genomics of
603 mediterranean oat (a. sativa) reveals high genetic diversity and three loci for
604 heading date. *Theoretical and Applied Genetics*.
605 Canales, F. J., Montilla-Bascón, G., Bekele, W. A., Howarth, C., Langdon, T.,
606 Rispail, N., Tinker, N., and Prats, E. (2021b). Data set from: Population
607 genomics of mediterranean oat (a. sativa) reveals high genetic diversity and three
608 loci for heading date. *Dryad, Dataset*.
609 Carlson, M. O., Montilla-Bascon, G., Hoekenga, O. A., Tinker, N. A., Poland,
610 J., Baseggio, M., Sorrells, M. E., Jannink, J.-L., Gore, M. A., and Yeats, T. H.
611 (2019). Multivariate genome-wide association analyses reveal the genetic basis
612 of seed fatty acid composition in oat (avena sativa l.). *G3: Genes, Genomes,*
613 *Genetics*, 9(9):2963–2975.

Chaffin, A. S., Huang, Y.-F., Smith, S., Bekele, W. A., Babiker, E., Gnanesh, B. N., Foresman, B. J., Blanchard, S. G., Jay, J. J., Reid, R. W., et al. (2016). A consensus map in cultivated hexaploid oat reveals conserved grass synteny with substantial subgenome rearrangement. *The plant genome*, 9(2):1–21.

Chen, L., Schenkel, F., Vinsky, M., Crews, D. H., and Li, C. (2013). Accuracy of predicting genomic breeding values for residual feed intake in angus and charolais beef cattle. *Journal of Animal Science*, 91:4669–4678.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323 – 329.

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. (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. *Genetics Selection Evolution*, 44(1):4.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345.

de Roos, A. P. W., Hayes, B. J., and Goddard, M. E. (2009). Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–1553.

Duhnen, A., Gras, A., Teyssèdre, S., Romestant, M., Claustres, B., Daydé, J., and Mangin, B. (2017). Genomic selection for yield and seed protein content in soybean: A study of breeding program data and assessment of prediction accuracy. *Crop Sci.*, 57:1–13.

Edwards, S. M., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., et al. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theoretical and Applied Genetics*, 132(7):1943–1952.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *The Plant Genome*, 4(3):250–255.

Esvelt Klos, K., Huang, Y.-F., Bekele, W. A., Obert, D. E., Babiker, E., Beattie, A. D., Bjørnstad, Å., Bonman, J. M., Carson, M. L., Chao, S., et al. (2016). Population genomics related to adaptation in elite oat germplasm. *The Plant Genome*, 9(2):1–12.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

FAO. (2017). *FAO statistical yearbook*. FAO.

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173(3):1761–1776.

Gianola, D. and van Kaam, J. B. C. H. M. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178(4):2289–2303.

Gorjanc, G., Jenko, J., Hearne, S. J., and Hickey, J. M. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics*, 17(1).

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., McMullen, M. D., Holland, J. B., Szalma, S. J., Wisser, R. J., and Yu, J. (2019). Optimal designs for genomic selection in hybrid crops. *Molecular Plant*, 12(3):390–401.

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., Xu, Z., Wang, D., and Gay, G. (2014). The impact of population structure on genomic prediction

663 in stratified populations. *Theoretical and Applied Genetics*, 127(3):749–762.
664 Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of
665 the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(1):186.
666 Haikka, H., Knurr, T., Manninen, O., Pietila, L., Isolahti, M., Teperi, E., Mantysaari,
667 E. A., and Stranden, I. (2020a). Genomic prediction of grain yield in commercial
668 Finnish oat (Avena sativa) and barley (Hordeum vulgare) breeding programmes.
669 *PLANT BREEDING*, 139(3):550–561.
670 Haikka, H., Manninen, O., Hautsalo, J., Pietila, L., Jalli, M., and Vetelainen, M.
671 (2020b). Genome-wide Association Study and Genomic Prediction for Fusarium
672 graminearum Resistance Traits in Nordic Oat (Avena sativa L.). *AGRONOMY-*
673 *BASEL*, 10(2).
674 Heslot, N., Yang, H., Sorrells, M. E., and Jannink, J. (2012). Genomic selection in
675 plant breeding: A comparison of models. *Crop Sci.*, 52:146–160.
676 Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M.,
677 Grondona, M., Zambelli, A., Windhausen, V. S., Mathews, K., et al. (2014).
678 Evaluation of genomic selection training population designs and genotyping
679 strategies in plant breeding programs using simulation. *Crop Science*, 54(4):1476–
680 1488.
681 Huang, Y.-F., Poland, J. A., Wight, C. P., Jackson, E. W., and Tinker, N. A.
682 (2014). Using genotyping-by-sequencing (gbs) for genomic discovery in cultivated
683 oat. *PLOS ONE*, 9(7):1–16.
684 Isidro, J., Akdemir, D., and Burke, J. (2016). Genomic selection. In William, A.,
685 Alain, B., and Maarten, V. G., editors, *The world wheat book: a history of wheat*
686 *breeding*, volume 3, chapter 32, pages 1001–1023. Lavoisier,Paris.
687 Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E.
688 (2015). Training set optimization under population structure in genomic selection.
689 *Theoretical and Applied Genetics*, 128(1):145–158.
690 Isidro-Sánchez, J., D'Arcy Cusack, K., Verheecke-Vaessen, C., Kahla, A., Bekele,
691 W., Doohan, F., Magan, N., and Medina, A. (2020a). Genome-wide association
692 mapping of fusarium langsethiae infection and mycotoxin accumulation in oat
693 (avena sativa l.). *The Plant Genome*, page e20023.
694 Isidro-Sánchez, J., Prats, E., Howarth, C., Langdon, T., and Montilla-Bascón, G.
695 (2020b). Genomic approaches for climate resilience breeding in oats. In *Genomic*
696 *Designing of Climate-Smart Cereal Crops*, pages 133–169. Springer.
697 Jiang, Y. and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics*,
698 201(2):759–768.
699 Karoui, S., Carabaño, M. J., Díaz, C., and Legarra, A. (2012). Joint genomic
700 evaluation of french dairy cattle breeds using multiple-trait models. *Genetics*
701 *Selection Evolution*, 44(1):39.
702 Kaufman, L. and Rousseeuw, P. (1987). *Clustering by Means of Medoids*. Delft
703 University of Technology : reports of the Faculty of Technical Mathematics and
704 Informatics. Faculty of Mathematics and Informatics.
705 Kebede, A. Z., Friesen-Enns, J., Gnanesh, B. N., Menzies, J. G., Fetch, J. W. M.,
706 Chong, J., Beattie, A. D., Paczos-Grzeda, E., and McCartney, C. A. (2019).
707 Mapping oat crown rust resistance gene pc45 confirms association with pckm.
708 *G3: Genes, Genomes, Genetics*, 9(2):505–511.
709 Laloë, D. (1993). Precision and information in linear models of genetic evaluation.
710 *Genetics Selection Evolution*, 25(6):557.

Laporte, F. and Mary-Huard, T. (2020). *MM4LMM: Inference of Linear Mixed Models Through MM Algorithm*. R package version 2.0.2.

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., Campo, L., Flament, P., Melchinger, A. E., Menz, M., Meyer, N., Moreau, L., Moreno-González, J., Ouzunova, M., Pausch, H., Ranc, N., Schipprack, W., Schönleben, M., Walter, H., Charcosset, A., and Schön, C.-C. (2014). Usefulness of multiparental populations of maize (zea mays l.) for genome-based prediction. *Genetics*, 198(1):3–16.

Lehermeier, C., Schön, C.-C., and de los Campos, G. (2015). Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics*, 201(1):323–337.

Lorenz, A. J. and Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop science*, 55(6):2657–2667.

Mangin, B., Rincent, R., Rabier, C.-E., Moreau, L., and Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of ethacc. *PLOS ONE*, 14(2):1–21.

Maughan, P. J., Lee, R., Walstead, R., Vickerstaff, R. J., Fogarty, M. C., Brouwer, C. R., Reid, R. R., Jay, J. J., Bekele, W. A., Jackson, E. W., Tinker, N. A., Langdon, T., Schlueter, J. A., and Jellen, E. N. (2019). Genomic insights from the first chromosome-scale assemblies of oat (avena spp.) diploid species. *BMC Biology*, 17(1).

Mellers, G., Mackay, I., Cowan, S., Griffiths, I., Martinez-Martin, P., Poland, J. A., Bekele, W., Tinker, N. A., Bentley, A. R., and Howarth, C. J. (2020). Implementing within-cross genomic prediction to reduce oat breeding costs. *The Plant Genome*, 13(1):e20004.

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

Montilla-Bascón, G., Rispail, N., Sánchez-Martín, J., Rubiales, D., Mur, L. A. J., Langdon, T., Howarth, C. J., and Prats, E. (2015). Genome-wide association study for crown rust (puccinia coronata f. sp. avenae) and powdery mildew (blumeria graminis f. sp. avenae) resistance in an oat (avena sativa) collection of commercial varieties and landraces. *Frontiers in Plant Science*, 6.

Montilla-Bascón, G., Sánchez-Martín, J., Rispail, N., Rubiales, D., Mur, L., Langdon, T., Griffiths, I., Howarth, C., and Prats, E. (2013). Genetic diversity and population structure among oat cultivars and landraces. *Plant molecular biology reporter*, 31(6):1305–1314.

Newell, M., Cook, D., Tinker, N., and Jannink, J.-L. (2011). Population structure and linkage disequilibrium in oat (avena sativa l.): implications for genome-wide association studies. *Theoretical and Applied Genetics*, 122(3):623–632.

Newell, M. A., Asoro, F. G., Scott, M. P., White, P. J., Beavis, W. D., and Jannink, J.-L. (2012). Genome-wide association study for oat (avena sativa l.) beta-glucan concentration using germplasm of worldwide origin. *Theoretical and Applied Genetics*, 125(8):1687–1696.

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3: Genes, Genomes, Genetics*, 8(9):2889–2899.

Olatoye, M. O., Clark, L. V., Labonte, N. R., Dong, H., Dwiyanti, M. S., Anzoua, K. G., Brummer, J. E., Ghimire, B. K., Dzyubenko, E., Dzyubenko, N., et al.

(2020). Training population optimization for genomic selection in miscanthus. *G3: Genes, Genomes, Genetics*, 10(7):2465–2476.

Olson, K. M., Van Raden, P. M., and Tooker, M. E. (2012). Multibreed genomic evaluations using purebred holsteins, jerseys, and brown swiss. *Journal of Dairy Science*, 95(9):5378–5383.

Ou, J.-H. and Liao, C.-T. (2019). Training set determination for genomic selection. *Theoretical and Applied Genetics*, 132(10):2781–2792.

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

Poland, J. A. and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, 5(3):92–102.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Pryce, J. E., Gredler, B., Bolormaa, S., Bowman, P. J., Egger-Danner, C., Fuerst, C., Emmerling, R., Solkner, J., Goddard, M. E., and Hayes, B. J. (2011). Short communication: Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science*, 94(5):2625–2630.

Pszczola, M., Strabel, T., Mulder, H., and Calus, M. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of dairy science*, 95(1):389–400.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics*, 130(11):2231–2247.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., and Moreau, L. (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(2):715–728.

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. *Theoretical and Applied Genetics*, 132(1):81–96.

Rispail, N., Montilla-Bascón, G., Sánchez-Martín, J., Flores, F., Howarth, C., Langdon, T., Rubiales, D., and Prats, E. (2018). Multi-environmental trials reveal genetic plasticity of oat agronomic traits associated with climate variable changes. *Frontiers in Plant Science*, 9:1358.

Roth, M., Muranty, H., Di Guardo, M., Guerra, W., Patocchi, A., and Costa, F. (2020). Genomic prediction of fruit texture and training population optimization towards the application of genomic selection in apple. *Horticulture research*, 7(1):1–14.

Sánchez-Martín, J., Rispail, N., Flores, F., Emeran, A. A., Sillero, J. C., Rubiales, D., and Prats, E. (2016). Higher rust resistance and similar yield of oat landraces versus cultivars under high temperature and drought. *Agronomy for Sustainable Development*, 37(1).

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., Mason, R. E., Babar, A., Harrison, S., Sutton, R., et al. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical usa winter wheat panel. *Theoretical and Applied Genetics*, 132(4):1247–

1261.

Sunstrum, F. G., Bekele, W. A., Wight, C. P., Yan, W., Chen, Y., and Tinker, N. A. (2019). A genetic linkage map in southern-by-spring oat identifies multiple quantitative trait loci for adaptation and rust resistance. *Plant breeding*, 138(1):82–94.

Sánchez-Martín, J., Rubiales, D., Flores, F., Emeran, A., Shtaya, M., Sillero, J., Allagui, M., and Prats, E. (2014). Adaptation of oat (avena sativa) cultivars to autumn sowings in mediterranean environments. *Field Crops Research*, 156:111 – 122.

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., Chabert-Martinello, M., Magnin-Robert, J.-B., Marget, P., Aubert, G., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Frontiers in plant science*, 6:941.

Technow, F., Burger, A., and Melchinger, A. E. (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes—Genomes—Genetics*, 3(2):197–203.

Tinker, N. A., Bekele, W. A., and Hattori, J. (2016). Haplotag: Software for haplotype-based genotyping-by-sequencing analysis. *G3: Genes, Genomes, Genetics*, 6(4):857–863.

Tinker, N. A., Chao, S., Lazo, G. R., Oliver, R. E., Huang, Y.-F., Poland, J. A., Jellen, E. N., Maughan, P. J., Kilian, A., and Jackson, E. W. (2014). A snp genotyping array for hexaploid oat. *The Plant Genome*, 7(3):1–8.

Tumino, G., Voorrips, R. E., Morcia, C., Ghizzoni, R., Germeier, C. U., Paulo, M.-J., Terzi, V., and Smulders, M. J. (2017). Genome-wide association analysis for lodging tolerance and plant height in a diverse european hexaploid oat collection. *Euphytica*, 213(8):163.

Tumino, G., Voorrips, R. E., Rizza, F., Badeck, F. W., Morcia, C., Ghizzoni, R., Germeier, C. U., Paulo, M.-J., Terzi, V., and Smulders, M. J. (2016). Population structure and genome-wide association analysis for frost tolerance in oat using continuous snp array signal intensity ratios. *Theoretical and Applied Genetics*, 129(9):1711–1724.

Welch, R. W. (2012). *The oat crop: production and utilization*. Springer Science & Business Media.

Winkler, L. R., Michael Bonman, J., Chao, S., Admassu Yimer, B., Bockelman, H., and Esvelt Klos, K. (2016). Population structure and genotype–phenotype associations in a collection of oat landraces and historic cultivars. *Frontiers in plant science*, 7:1077.

Yan, H., Bekele, W. A., Wight, C. P., Peng, Y., Langdon, T., Latta, R. G., Fu, Y.-B., Diederichsen, A., Howarth, C. J., Jellen, E. N., et al. (2016). High-density marker profiling confirms ancestral genomes of avena species and identifies d-genome chromosomes of hexaploid oat. *Theoretical and Applied Genetics*, 129(11):2133–2149.

Yan, H., Zhou, P., Peng, Y., Bekele, W. A., Ren, C., Tinker, N. A., and Peng, Y. (2020). Genetic diversity and genome-wide association analysis in chinese hulless oat germplasm. *Theoretical and Applied Genetics*, pages 1–16.

Zhou, L., Ding, X., Zhang, Q., Wang, Y., Lund, M. S., and Su, G. (2013). Consistency of linkage disequilibrium between chinese and nordic holsteins and genomic prediction for chinese holsteins using a joint reference population. *Genetics Selection Evolution*, 45(1):7.

Supplementary Materials

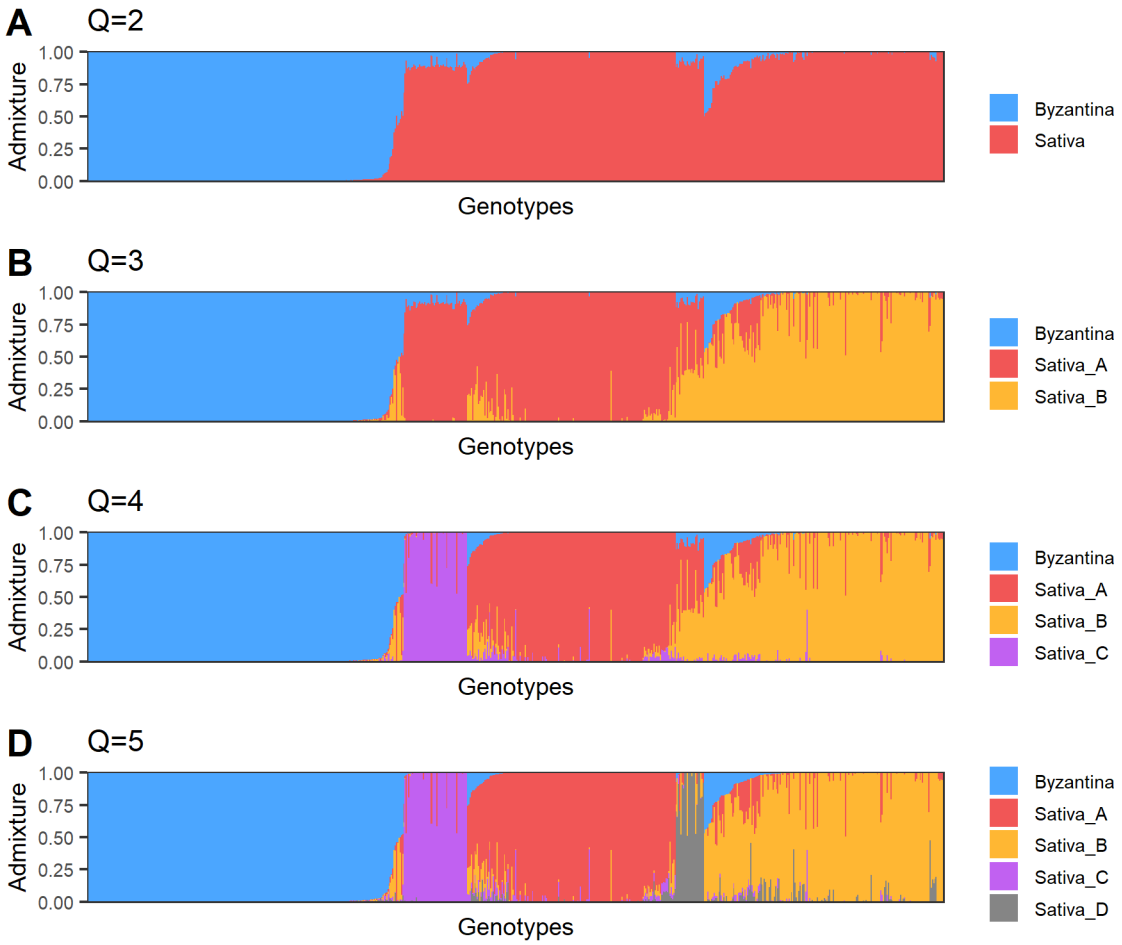
Genomic prediction and training set optimization in a structured Mediterranean oat population

Simon Rio[1*], Luis Gallego-Sánchez[2*], Gracia Montilla-Bascón[2], Francisco J. Canales[2], Julio Isidro y Sánchez[1], and Elena Prats[2]

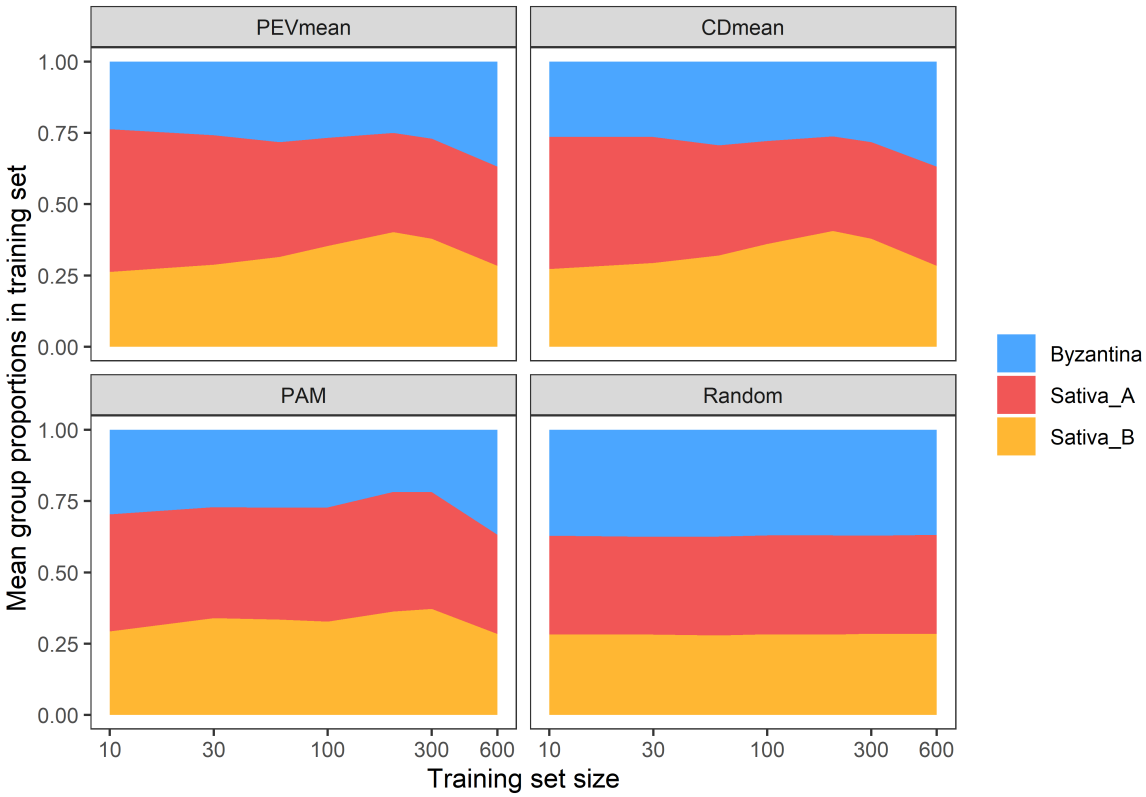
[1]Centro de Biotecnologia y Genómica de Plantas (CBGP, UPM-INIA) Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnologia Agraria y Alimentaria (INIA) Campus de Montegancedo-UPM 28223-Pozuelo de Alarcón, (Madrid), Spain

[2]Institute for Sustainable Agriculture, Spanish Research Council (CSIC), Córdoba, Spain

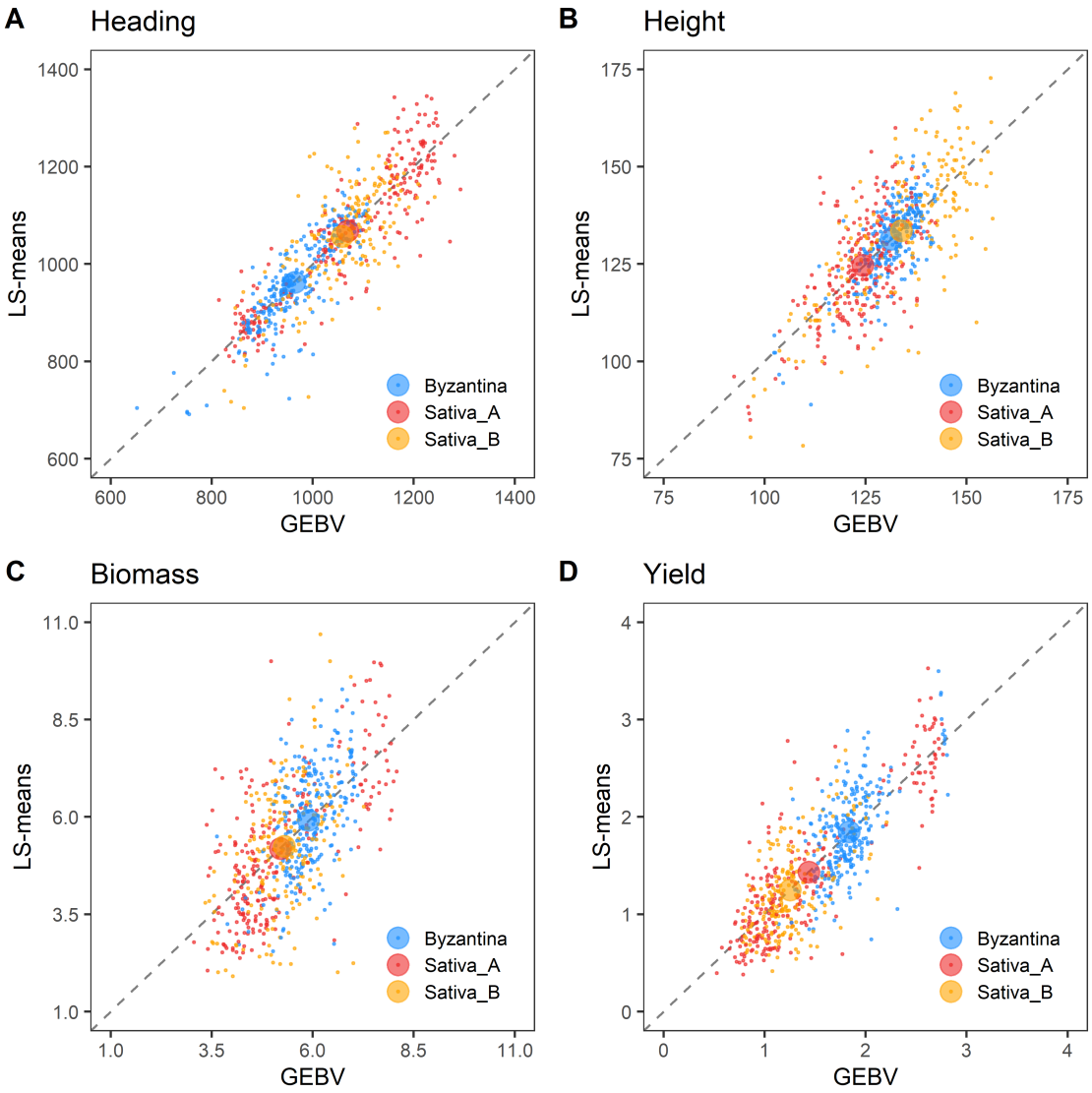
[*]These authors contributed equally to this study



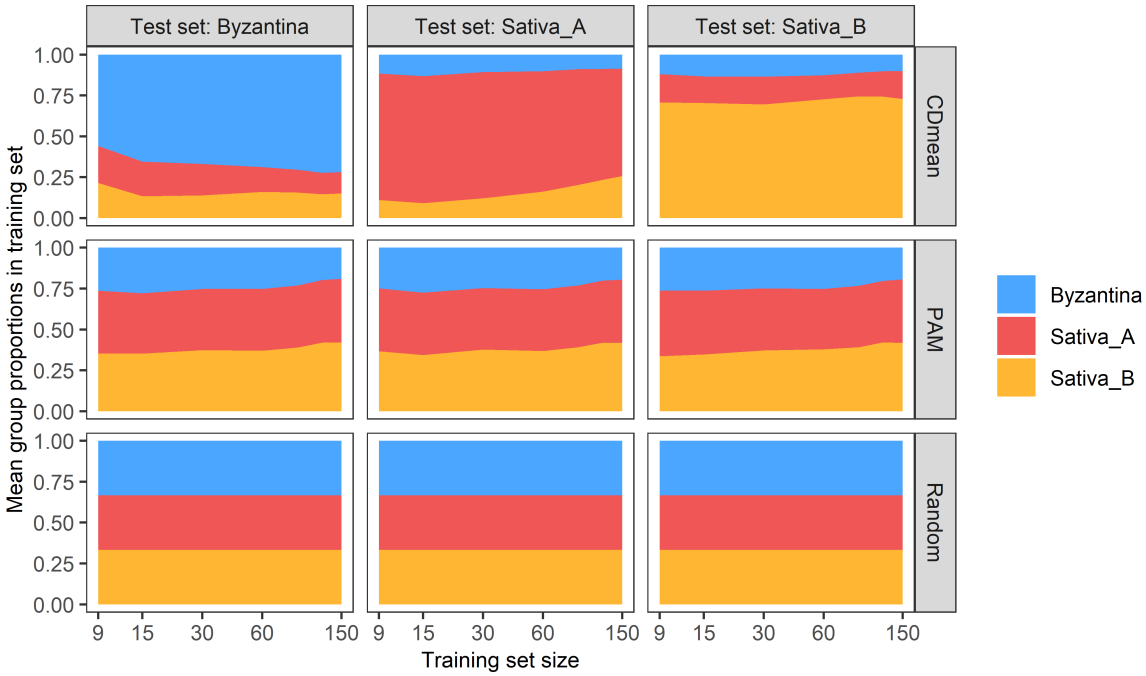
Supplementary Figure S1. Admixture barplots showing the admixture proportions of each individual and obtained using STRUCTURE for different numbers of genetic groups: (**A**) $Q = 2$, (**B**) $Q = 3$, (**C**) $Q = 4$, and (**D**) $Q = 5$.



Supplementary Figure S2. Mean group proportions in the training set (TRS) according to TRS size obtained using different optimization methods (over 30 random test sets (TSs)): PEVmean, CDmean and PAM, along with random sampling as a benchmark (x 30 replicates for each TS)



Supplementary Figure S3. Plots of genomic estimated breeding values (GEBVs) against LS-means obtained by leave-one-out (LOO) cross-validation using GBLUP for (**A**) Heading (in growing degree-days (GDD), (**B**) Height (in cm), (**C**) Biomass (in t/ha) and (**D**) Yield (in t/ha). Each dot represents one individual and was colored according to its genetic group. Big dots represent the mean of GEBVs and LS-means for each group.



Supplementary Figure S4. Mean group proportions in the training set (TRS) according to TRS size obtained using different optimization methods (over 30 random test sets (TSs)): CDmean and PAM, along with random sampling with equal group contributions in the TRS as a benchmark (x 30 replicates for each TS). Each candidate set consisted in a of 450 individuals, including 150 of each of the three genetic groups.