




Genomic prediction in family bulks using different traits and cross-validations in pine

Esteban F. Rios ^{1,*}, Mario H. M. L. Andrade,¹ Marcio F. R. Resende Jr.,² Matias Kirst,³ Marcos D. V. de Resende,⁴ Janeo E. de Almeida Filho ⁵, Salvador A. Gezan,⁶ and Patricio Munoz ²

¹Agronomy Department, University of Florida, Gainesville, FL 32611, USA

²Horticultural Sciences Department, University of Florida, Gainesville, FL 32611, USA

³School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611, USA

⁴EMBRAPA Café/Department of Statistics, Federal University of Viçosa, Avenida PH Rolfs S/N, Viçosa 36570-000, Brazil

⁵Bayer Crop Science, Estrada da Internadinha, 2000, Coxilha-RS 99145-000, Brazil

⁶VSN International Ltd, Hemel Hempstead HP2 4TP, UK

*Corresponding author: Agronomy Department, University of Florida, 2005 SW 23rd Street, Building 350 Off 5, Gainesville, FL 32608, USA.
Email: estebanrios@ufl.edu

Abstract

Genomic prediction integrates statistical, genomic, and computational tools to improve the estimation of breeding values and increase genetic gain. Due to the broad diversity in mating systems, breeding schemes, propagation methods, and unit of selection, no universal genomic prediction approach can be applied in all crops. In a genome-wide family prediction (GWFP) approach, the family is the basic unit of selection. We tested GWFP in two loblolly pine (*Pinus taeda* L.) datasets: a breeding population composed of 63 full-sib families (5–20 individuals per family), and a simulated population with the same pedigree structure. In both populations, phenotypic and genomic data was pooled at the family level *in silico*. Marker effects were estimated to compute genomic estimated breeding values (GEBV) at the individual and family (GWFP) levels. Less than six individuals per family produced inaccurate estimates of family phenotypic performance and allele frequency. Tested across different scenarios, GWFP predictive ability was higher than those for GEBV in both populations. Validation sets composed of families with similar phenotypic mean and variance as the training population yielded predictions consistently higher and more accurate than other validation sets. Results revealed potential for applying GWFP in breeding programs whose selection unit are family, and for systems where family can serve as training sets. The GWFP approach is well suited for crops that are routinely genotyped and phenotyped at the plot-level, but it can be extended to other breeding programs. Higher predictive ability obtained with GWFP would motivate the application of genomic prediction in these situations.

Keywords: family selection; training population; predictive ability; genomic prediction

Introduction

Genomic (Elshire *et al.* 2011), statistical (Meuwissen *et al.* 2001; Gianola *et al.* 2009), and computational advances have allowed significant increases in genetic gain by applying genomic prediction in breeding programs across several species (*e.g.*, Hayes *et al.* 2009; Fè *et al.* 2015, 2016; Gezan *et al.* 2017; Amadeu *et al.* 2020; de Bem Oliveira *et al.* 2020). Taking advantage of the ever-reducing cost of molecular markers (Wetterstrand, 2020), the concept of genomic prediction was derived (Meuwissen *et al.* 2001) as an alternative method to marker-assisted selection. Genomic prediction utilizes a dense panel of molecular markers covering the whole genome to predict genomic estimated breeding values (GEBV) of individuals with no phenotypic records (Meuwissen *et al.* 2001). Traditional genomic prediction pipelines involve developing a training set, for which available genotypic and phenotypic data is fitted to build a prediction model. This model is later used to predict GEBV of selection candidates in a validation set, composed of individuals that are genotyped but not phenotyped.

Cross-validation schemes are implemented taking sub-samples from the training set to calibrate the model and then fit the model into the remaining part of the training set to estimate and evaluate its predictive ability, *i.e.*, the correlation between GEBVs and phenotypic values (Pérez-Cabal *et al.* 2012).

Genomic prediction has been quickly adopted in animal breeding (Hayes *et al.* 2009) due to readily accessible genomic data, large reference populations with accurate pedigree records, and the impossibility of phenotyping sex-linked traits (Stock and Reents 2013). In dairy cattle, genomic prediction can double the genetic gain compared with selection based on progeny test (Xu *et al.* 2020). On the contrary, the application of genomic prediction in plants has been lagging behind due to less accessible high-throughput genotyping methods, lack of accurate pedigree records, and the wide range of variation in life cycle, ploidy level, and mating systems found in plants (Hough *et al.* 2013). All these plant-specific characteristics are key factors affecting predictive ability in genomic prediction due to their influence in breeding

Received: March 09, 2021. Accepted: July 02, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

methods, effective population size, population structure, and linkage disequilibrium (Lin et al. 2014). Pioneer studies implementing genomic prediction in plants were performed in major crop species with traditional hybrid selection such as maize (Combs and Bernardo 2013; Massman et al. 2013) and trees (Kumar et al. 2012; Resende et al. 2012), or variety selection in self-pollinating species (Poland et al. 2012). Genomic prediction showed to be a powerful tool to achieve higher genetic gain in plant breeding in many other species (Crossa et al. 2017; Lara et al. 2019; de Bem Oliveira et al. 2020; Esfandyari et al. 2020). Large commercial breeding companies have been applying genomic prediction; however, the success of the process depends strongly on the species and the breeding program scheme (Voss-Fels et al. 2019; Xu et al. 2020).

Several species are bred as populations of large full or half-sib families, and commercially used as populations of different levels of relationship (i.e., synthetic cultivars) as in some forage species, such as alfalfa (*Medicago sativa* L.; Annicchiarico et al. 2015; Biazzi et al. 2017) and ryegrass (*Lolium perenne* L.; Fè et al. 2016; Cericola et al. 2018). In those species, the family (full or half-sibs) is the basic unit for phenotyping (e.g., plot-level measurement for yield rather than plant level) and selection. Thus, due to the mating system nature (allogamy), individual plants are of limited interest because commercial varieties represent a homogenous population composed of heterozygous individuals (Poehlman 1987). Also, it is not straightforward to link phenotypic data collected on individual spaced-plants to plot-based swards in crops such as forage and turfgrass, which are mostly allogamous (Poehlman 1987), and single-plant performance has been shown to poorly predict plot-based data (Wang et al. 2016). Therefore, the application of genome-wide family prediction (GWFP) would be advantageous for traits that are phenotyped using family pools in swards or plots. The phenotypic data collection at the plot level could be extended to other organisms grown and evaluated in families, such as turfgrasses (*L. perenne* L.), forages (*M. sativa* L.), sugarcane (*Saccharum officinarum* L.), cassava (*Manihot esculenta* L.), honey bees, and to aquaculture species such as shrimp (*Litopenaeus vannamei*; Barbosa et al. 2012; Wang et al. 2017; Jia et al. 2018; Pembleton et al. 2018; Brascamp and Bijma 2019; Torres et al. 2019). The application of GWFP has already been reported for crops that are bred and farmed as family pools, such as cross-pollinated forage species (Annicchiarico et al. 2015; Fè et al. 2015, 2016; Biazzi et al. 2017; Cericola et al. 2018; Guo et al. 2018; Jia et al. 2018).

The GWFP approach considers family-pools as the measurement unit. Here, both allele frequencies and phenotypic records are expressed as a single-average record of a given family. Therefore, the additive genetic variance in full-sib families is half of the additive variance between individuals. Full-sibs share the same parents, hence the mean genotypic value of a full-sib family is equal to the mean breeding value of the two parents: $\frac{1}{4}(V_a + V_a) = \frac{1}{2}V_a$. This $\frac{1}{2}V_a$ variance represents the additive genetic variance among full-sib families, whereas the other $\frac{1}{2}V_a$ is the variance within a family, i.e., the variance between individuals (i.e., only 50% of the genetic variation is exploited in GWFP; Falconer and Mackay 1996). As a result, higher predictive ability was reported in family pools when compared with GEBV (Ashraf et al. 2014). Despite the initial efforts to test the predictive ability of GWFP using empirical data, there is a need to explore further implementation of GWFP in breeding schemes. As a first aspect, it is essential to compare the predictive ability of GEBV vs GWFP models, and to develop strategies to combine both approaches. For this, datasets that contain family structures but genotyped

and phenotyped at the single-plant level are ideal. Another aspect is the understanding of the influence that family/pool size and phenotypic variances in training/validation sets have in the predictive ability for various traits.

In order to evaluate these aspects, two loblolly pine (*Pinus taeda* L.) populations were studied: (1) an observed breeding population composed of 63 families (CCLONES_real), and (2) a simulated population that reproduced the same pedigree as CCLONES_real. The objectives of this study are: (i) to identify the minimum number of individuals per family required to calculate allele frequency and phenotypic mean values with reasonable accuracy; (ii) to investigate the effect of contrasting phenotypic mean and variance between training and validation sets on predictive ability; and (iii) to assess the predictive ability of GEBV and GWFP. Loblolly pine is not normally bred in family pools, but existing real and simulated datasets were used to compare GEBV and GWFP approaches.

Materials and methods

Loblolly pine real population data

The phenotypic data from the loblolly pine (*P. taeda* L.) population known as “comparing clonal lines on experimental sites” (CCLONES_real), which have previously been used for predicting the performance of individual trees (Resende et al. 2012), was used to assess the efficiency of the GWFP. In this study, GWFP was tested by pooling individual trees belonging to the same full-sib family. The population is composed of 923 individuals from 70 full-sib families obtained by crossing 32 parents in a circular diallel mating design with additional off-diagonal crosses (Baltunis et al. 2007). The number of individuals per family ranged from 1 to 20, with an average of 13 trees per family (standard deviation = 5). In this study, families with less than five individuals were removed, and 63 full-sib families were used for analyses. Data collection was described in detail in Resende et al. (2012) and Munoz et al. (2014). In summary, all 923 genotypes from CCLONES_real was phenotypically characterized in three replicated studies and was genotyped using an Illumina Infinium assay (Illumina, San Diego, CA, USA; Eckert et al. 2010) with 7216 SNPs, each representing a unique pine EST contig. In this study, four traits representing growth, quality, and diseases were selected based on their narrow-sense heritability and genetic architecture as reported by Resende et al. (2012). These correspond to: (1) lignin concentration (Lignin) ($h^2 = 0.11$, polygenic trait), (2) tree stiffness (Stiffness) at year 4 (km^2/s^2) ($h^2 = 0.37$, polygenic trait), (3) rust susceptibility (Rust) caused by *Cronartium quercuum* Berk. Miyable ex Shirai f. sp. *Fusifforme* ($h^2 = 0.21$, oligogenic trait), and (4) diameter at breast height (Diameter) at year 6 (cm) ($h^2 = 0.31$, polygenic trait).

Simulated population

A simulated population (CCLONES_sim) exhibiting similar genetic properties as CCLONES_real was also considered in this study, aiming to assess the efficiency of GWFP for two different traits and to predict the performance of the next generation. The description for the simulation, and the results for genomic prediction approaches using individual trees (GEBV) were previously reported for this synthetic population (de Almeida Filho et al. 2016, 2019). In summary, the base population was created ($G_0 = 1000$ diploid individuals) by randomly sampling 2000 haplotypes from a population with an effective size of $N_e = 10,000$ (Johnson et al. 2001) and a mutation rate of 2.5×10^{-8} . Then, the 10% highest phenotypic values from G_0 were selected and

randomly mated to generate the first breeding generation (G1). From G1, 42 individuals were selected and used in a circular diallel mating design that reproduced the pedigree as in CCLONES_real (G2), comprised of 923 individuals and 71 full-sib families. However, only 63 families, with more than five individuals, were used in this study. Subsequently, 42 individuals were selected from G2 and used in crosses to the next generation (G3, CCLONES_sim_prog), a population composed of 1176 individuals and 71 families. Only the 63 families with more than 5 individuals were used for analyses.

The simulated genome had 12 chromosomes, each with 100 cM, and 10,000 polymorphic loci were randomly selected to represent the entire genome, and only the scenario exhibiting an absence of dominance ($d^2 = 0.0$) and $h^2 = 0.25$ were used for analyses in this study. Two traits with different genetic architectures were simulated: (1) oligogenic: 30 QTL were sampled from a gamma distribution with rate 1.66 and shape 0.4, with positive or negative QTL effects (Meuwissen et al. 2001), and (2) polygenic: 1000 QTL were used, and their additive effects were sampled from a standard normal distribution (Hickey and Gorjanc 2012). The simulations were run using Macs (Chen et al. 2009) and in the software R using scripts developed by the authors.

Pooling phenotypic and genotypic data at the family level

In both populations, phenotypic and genotypic data were pooled at the family level *in silico*. We assumed that the family phenotype was the average of all individuals in a family. Hence, the phenotypic value for each individual was pooled at the family level *in silico* by calculating the family mean, without considering the experimental design. Therefore, the average phenotypic value by family was used as the response for all analyses.

In the case of the genomic data, the allele frequency (p) was calculated for each SNP per family, considering the reference allele (A) as follows:

$$p_{ij} = (2n_{AA_{ij}} + n_{Aa_{ij}}) / 2N_{ij},$$

where p_{ij} refers to the allele frequency for SNP i in the j family; $n_{AA_{ij}}$ and $n_{Aa_{ij}}$ are number of individuals with genotype AA and Aa respectively for SNP i in the family j ; N_{ij} are number of individuals in family j with non-missing genotype data for SNP i . Missing values for allele frequency were imputed at the family level using the average allele frequency for that given SNP across families. Markers were excluded from analyses when more than 50% of the families exhibited missing values, and SNPs were not removed based on minor allele frequency. A total of 4740 polymorphic SNPs (CCLONES_real) and an average of 5000 polymorphic SNPs for CCLONES_sim and CCLONES_sim_prog (average across simulated replicates) were used in the analyses.

Minimum number of individuals per family to estimate allele frequency and family phenotypic mean

A total of 10 families from CCLONES_real and CCLONES_sim with at least 15 individuals were selected to evaluate the minimum number of individuals required to estimate allele frequency and phenotypic family means with the most reasonable accuracy. Families were specifically selected to represent segregation ratios (1:1 and 1:2:1) for 10 SNPs. Allele frequencies per family and family phenotypic means were calculated varying the number of individuals per family from one to 15. These values were used to

compute the squared deviations between the mean value obtained with i number of individuals ($i=1-15$) and the mean value obtained with the entire family (15 individuals), under the assumption that 15 individuals per family provide accurate estimates of allele frequencies and phenotypic mean in our families. This assumption can be validated using the concept of genetic representativeness, given by the effective population size (N_e) (Vencovsky and Crossa 2003). The estimator of the N_e within a full sib family is given by $N_e = [2n/(n+1)]$ (Resende and Barbosa 2006). The maximum (when n goes to infinite) N_e within a full sib family is 2. With n equal to 15 individuals the N_e is 1.88, which is 94% of this maximum of 2.

Statistical methods for genomic prediction

Marker effects were estimated at the individual (GEBV) and family (GWFP) levels with two distinct whole-genome regression approaches using the package BGLR (Perez and de los Campos, 2014) in R (R Development Core Team 2018): (1) Bayes B which considers that markers have heterogeneous variances, i.e., many loci with no genetic variance and a few loci explain a large portion of the genetic variation (Meuwissen et al. 2001; Pérez and de Los Campos 2014); and (2) Bayes RR a Bayesian method that assumes common variance across all loci; therefore, SNPs with the same allele frequency explain the same proportion of variance and have the same shrinkage effect (Gianola, 2013; Pérez and de Los Campos 2014).

In total, 20,000 Markov chain Monte Carlo iterations were used, of which the first 5000 were discarded as burn-in, and every third sample was kept for parameter estimation. We fitted the following model for individual and family models:

$$\mathbf{y} = 1\mu + \mathbf{Z}\mathbf{m} + \mathbf{e},$$

where \mathbf{y} is the vector of the averaged phenotype by family in the case of GWFP and by individual in the multiple clones in the case of GEBV, μ is the overall mean fitted as a fixed effect, \mathbf{m} is the vector of random marker effects, and \mathbf{e} is the vector of random error effects, $\mathbf{1}$ is a vector of ones, and \mathbf{Z} is the incidence matrix indicating allele frequencies in the case of GWFP (ranging from 0 to 1), and marker dosage (0, 1, and 2) for GEBV.

After fitting the model described above for each trait, the GWFP and GEBV of family/individual j (\hat{g}_j) were obtained using the following expression:

$$\hat{g}_j = \sum_1^p Z_{ij} \hat{m}_i,$$

where Z_{ij} is the allele frequency/marker dosage of the i th marker on family/individual j , and p is the total number of markers, and \hat{m}_i is the estimated effect of i th SNP.

Cross-validation schemes

The prediction models for GEBV and GWFP were validated using 10-fold cross-validation and leave-one-out approaches, for both populations and all traits. For the 10-fold cross-validation, data was randomly partitioned into ten subsets, and training set populations were created with 90% of the families/individuals, whereas the remaining 10% of families/individuals were used as validation set. This scheme was repeated until the ten subsets were used as validation set. In the leave-one-out approach, models were constructed using $N_T - 1$ families (where N_T is the total number of families) in the training set. The validation set was the

single family not included in the training group. This scheme was repeated N_T times until all 63 families were used as the training set.

Each time the models were fitted using a different validation set, the model's predictive ability was estimated calculating a Pearson's correlation between the observed/simulated phenotypes and the GWFP/GBV estimates for the families/individuals included in the validation set.

Creating training/validation sets using contrasting phenotypes

To assess the effect that the validation set structure has in the predictive ability of the models, both populations were divided in three different phenotypic classes for each trait: the smallest 10%, the largest 10%, and values between both extremes. Five validation sets were created for each trait using these phenotypic classes: (1) Low: 10% families with the lowest phenotypic values; (2) High: 10% families having the highest values; (3) Low+High: combining four families from Low and three families from High; (4) Middle: seven families showing phenotypes around the population mean, (5) Combined: two families from Low, two families from High, and three families from Middle. For the populations Low+High (3), Middle (4), and Combined (5), three replicates were created by taking random samples from each phenotypic class. The other 56 families were used as training sets to build prediction models.

Split-families as training/validation sets

Two scenarios were created to explore the ability of the GWFP models to predict the performance of individuals and family pools. All families with more than 10 individuals (59 families in total) were randomly split into 2 equivalent size groups. For one group of individuals phenotypic and genotypic data were pooled at the family level and used as the training set for GWFP models. The other group of individuals was used as the validation set based on two approaches: (1) predicting the performance of individuals trees not included in the training set (GWFP_Fam_Ind), and (2) pooling individuals at the family level to predict performance of families composed of individuals not included in the training set (GWFP_Fam_Fam).

Prediction in the following generation using GBV and GWFP in the simulated population

The genomic prediction models were developed by using the G2 CCLONES_sim population as the training set. These training models were used and validated in the G3 generation using GBV and GWFP, and models were assessed by calculating predictive ability and prediction accuracy. Predicted ability was estimated by calculating a Pearson's correlation between the phenotypic values and the estimated breeding values, and prediction accuracy was estimated by calculating a Pearson's correlation between the real breeding value and the estimated breeding value.

Results

Minimum number of individuals per family to estimate allele frequency and family phenotypic mean

The minimum number of individuals per family was calculated assessing allele frequency and phenotypic mean deviations using families with at least 15 individuals. For genotypic and phenotypic data, the lowest number of individuals needed to accurately estimate allele frequency and family means was six (Figure 1).

Allele frequency deviations (Figure 1, A–D) and mean phenotypic deviations (Figure 1, E and F) indicated that families with less than six individuals were not providing accurate estimates of the family's genotypic and phenotypic means in both populations. We assumed that the observed values based on 15 individuals per family provides with a reasonable estimation of allele frequency and phenotypic mean for a diploid species. Therefore, all 63 families with six or more individuals were used for further analyses in this study. Both populations showed similar trends for the genotypic and phenotypic estimates (Figure 1). The average allele frequency deviations were lower for SNPs exhibiting a 1:1 ratio in both populations (Figure 1, A and B), compared with SNPs segregating into a 1:2:1 ratio (Figure 1, C and D). For phenotypic data, CCLONES_sim showed slightly smaller deviations, especially for a lower number of individuals (Figure 1F), compared with CCLONES_real for the trait diameter (Figure 1E). Other traits in CCLONES_real exhibited a similar behavior (data not shown).

Predictive ability of statistical methods for genomic prediction and for different cross-validation schemes

Two Bayesian statistical methods (*Bayes B* and *Bayes RR*) and two cross-validation approaches were used to test the predictive ability of GWFP in four traits measured in CCLONES_real (Figure 2). Both statistical methods yielded high and similar predictive abilities for the four traits (Figure 2, A and B). However, standard errors for predictive ability were larger with the leave-one-out approach (Figure 2, A and B). Additionally, GWFP predictive abilities obtained with the leave-one-out approach were slightly lower than for the 10-fold cross-validation scheme (except for trait Stiffness) (Figure 2, A and B). Therefore, the 10-fold cross validation approach was selected to perform further analyses.

Predictive ability of GWFP using training/validation sets with contrasting phenotypes

The effect of phenotypic data in the predictive ability of GWFP was explored by creating five validation sets using contrasting sets of phenotypic data between training set and validation set (Figure 3A). The predictive ability for GWFP for all traits were least accurate and had larger standard errors when the validation set was composed of families exhibiting small and large phenotypic values (bottom and top classes; Figure 3B). When validation sets were composed of families exhibiting phenotypes corresponding to the middle class, predictive ability increased for all traits, but standard errors were still large (Figure 3B). As expected, there was an increase in predictive ability and a large reduction in standard errors when validation sets were composed of families showing similar phenotypic mean and variance to the training set, corresponding to the classes "Low+High" and "Combined" (Figure 3B).

Predictive ability of GBV and GWFP

Predictive ability obtained with *Bayes B* using different methods and schemes (Table 1) is presented in Figure 4 for the 63 families from both populations. The traditional genomic prediction approach with individuals in the training set and validation set (GBV) was contrasted with predictive ability obtained with the family-based (GWFP) method following a 10-fold cross validation scheme. The scenarios GWFP_Fam_Ind and GWFP_Fam_Fam were run only once because CCLONES (real and simulated) had a limited number of individuals per family.

Predictive ability was always greater for GWFP methods in both populations and all traits, except for the scenario

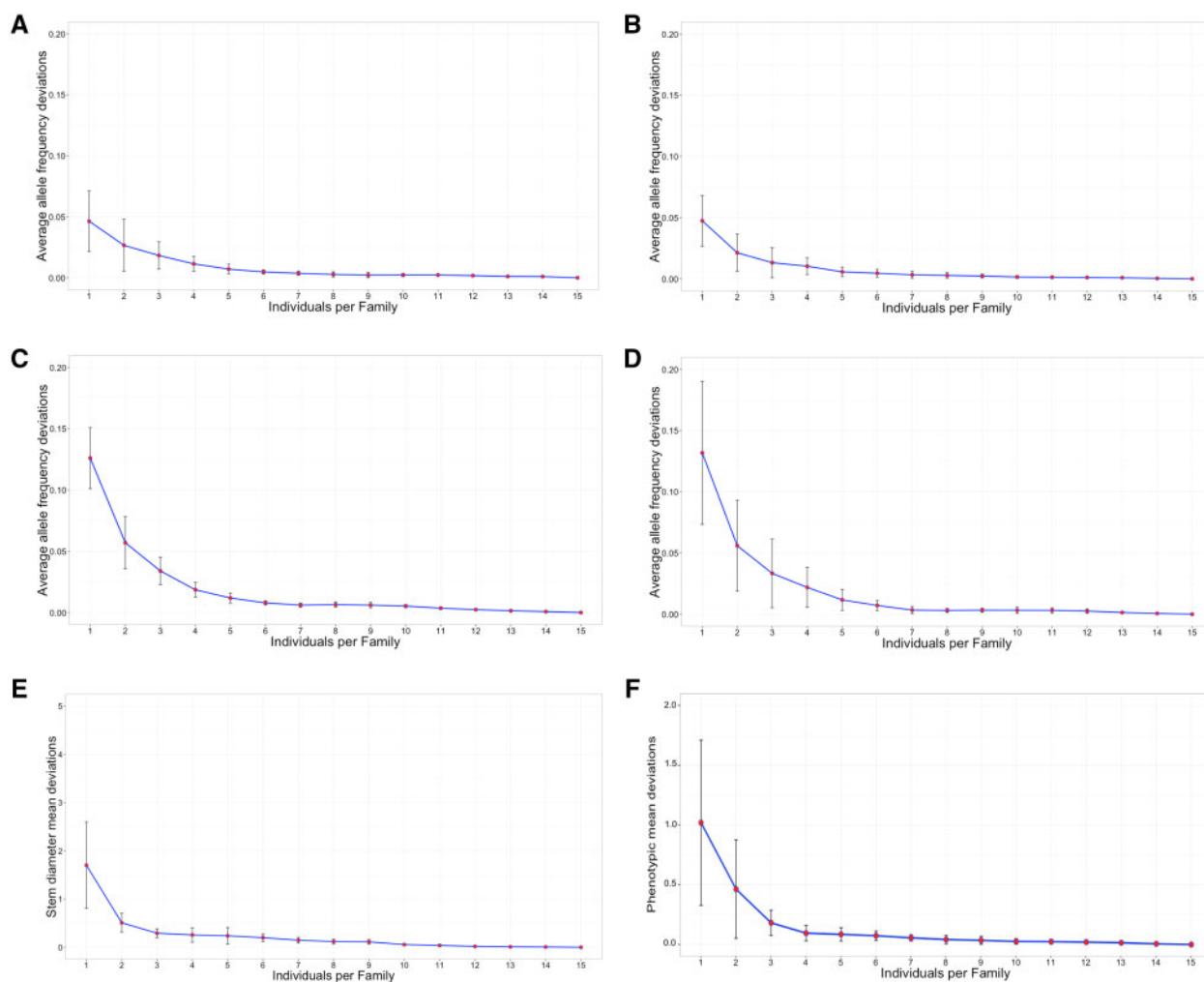


Figure 1 Average allele frequency deviation (A–D) and family mean phenotypic deviation (E and F) in CCLONES_real (real breeding population composed of 63 families) (A, C, and E) and CCLONES_sim (simulated breeding population exhibiting similar genetic properties of CCLONES_real) (B, D, and F) calculated by increasing the number of individuals from 1 to 15. Five families exhibiting genotypic segregation ratios 1:1 (A and B) and 1:2:1 (C and D) for single nucleotide polymorphisms were included in the analysis. The CCLONES_real phenotypic deviation is for the trait stem diameter (E).

GWFP_Fam_Ind that showed similar or lower accuracy than GEBV for most traits (Figure 4). Additionally, predictive ability was greater for traits with higher heritability (Figure 4). Specifically, GWFP provided predictive abilities at least 40% greater than traditional GEBV for most of the traits in both populations. Moreover, GWFP_Fam_Fam exhibited similar or greater predictive ability than GWFP for most traits in both populations, except for rust (Figure 4). Both sets of traits from the simulated CCLONES population exhibited very similar accuracies for all schemes (Figure 4).

Predictive ability and accuracy of GEBV and GWFP in the following generation

Accuracy and predictive ability of GEBV and GWFP were obtained with the prediction models built with the CCLONES_sim (G2) population as the training set, and models were validated in the following generation (G3). The GEBV showed higher accuracy than GWFP for the oligogenic trait, and similar accuracy for the polygenic trait (Figure 5). Predictive ability for the oligogenic and polygenic traits were higher for GWFP (Figure 5). Additionally, greater predictive ability and accuracy were observed for the oligogenic

trait, and the difference between accuracy and predictive ability was greater for the oligogenic trait (Figure 5).

Discussion

We quantified the predictive ability of GWFP in real and simulated loblolly pine breeding populations for different traits and cross-validation approaches. Moderate to low predictive ability values were obtained with the traditional genomic prediction approach, as previously reported for both populations, using individual trees as the basic phenotypic and genotypic unit (Resende et al. 2012; de Almeida Filho et al. 2016). In general, GWFP outperformed GEBV in the predictive ability for most traits; including the predictive ability for the oligogenic and polygenic traits in CCLONES_sim when using the following generation (G3) as the validation set.

Effect of family size in genomic prediction

The size and structure of the training population affects the accuracy of genomic prediction models (VanRaden et al. 2009; Daetwyler et al. 2010; Habier et al. 2010; Grattapaglia and Resende 2011; Edwards et al. 2019; de Bem Oliveira et al. 2020). In our study,

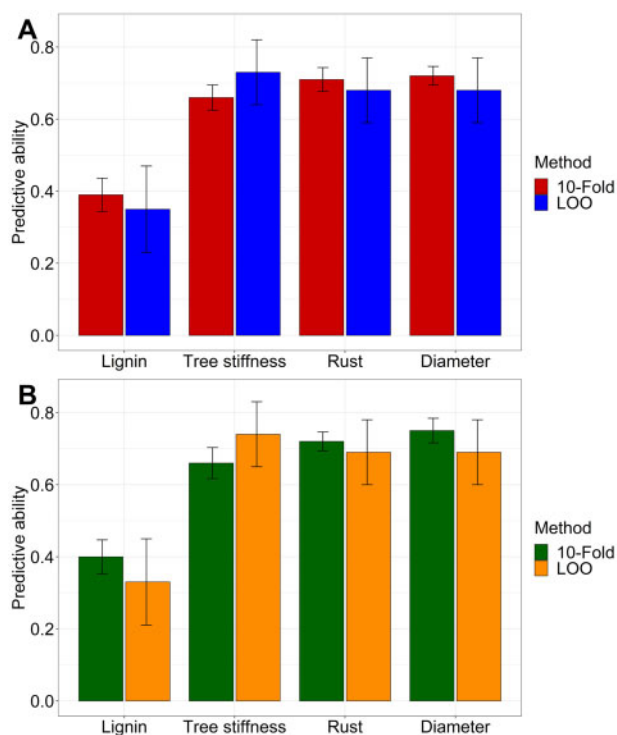


Figure 2 Average predictive ability using family pools (GWFP) in four traits in the loblolly pine breeding population CCLONES obtained with 10-fold and leave-one-out cross-validation schemes using Bayes B (A) and Bayes RR (B).

the size of the training set refers to the number of families and the number of individuals within a family. The number of families was fixed and limited to 70 families, so we did not focus on studying the effect of a variable number of families. However, the minimum number of individuals per family to obtain reasonable accurate estimates of family allele frequency and family phenotypic mean was found to be six. When studying the effect of size and composition of training population in blueberry (*Vaccinium* spp.), de Bem Oliveira et al. (2020) found a high predictive ability using six individuals per family for some traits. Thus, in their study family variance was accurately represented with six individuals per family in this autotetraploid species. Using the estimator of the N_e within a full sib family, given by $N_e = [2n/(n+1)]$ (Resende and Barbosa 2006), the maximum (when n goes to infinite) N_e within a full sib family is 2. With n equal to 6 individuals the N_e is 1.71, which is 86% of the maximum 2. So, $n=6$ appears adequate to represent genetically a full-sib family, corroborating our results.

The effect of number of individuals within families on accuracy of genomic prediction models was also demonstrated in perennial ryegrass (Pembleton et al. 2016, 2018). The authors stated that 48–60 individuals per population are necessary to accurately represent the genetic diversity within a ryegrass population. As an allogamous species, multiple parents are used to create synthetic populations in perennial ryegrass; hence, multiple individuals with a high number of loci in heterozygosity are contributing to the variation in the synthetic population. Perennial ryegrass is commonly bred using families and GWFP has been exploited in the species for various traits (Fè et al. 2015, 2016; Cericola et al. 2018; Guo et al. 2018).

Simulation studies with variable numbers of families and individuals per family would help identify the optimum

training population sizes for GWFP. Generally, a larger training population (more families in the training population) yield higher accuracy (Voss-Fels et al. 2019; de Bem Oliveira et al. 2020), but this is associated with higher costs. Therefore, the definition of the optimum number of families, and number of individuals per family are a crucial point for the genomic prediction process. Fé et al. (2015) studied the effect of the number of families in the accuracy of genomic prediction for heading date in ryegrass; the authors found high accuracies with a low number of families (<100). The authors showed that increasing the number of families to 500 leads to higher accuracy, and more than 500 families did not yield to significant improvement.

Efficiency of statistical methods and cross-validation schemes

Models considering different Bayesian methods were similar in predicting GEBV in traits measured in the real breeding population and the simulated population in this study. Resende et al. (2012), reported a slightly greater predictive ability in the real population for rust incidence with Bayesian methods over RR-BLUP, because fewer genes with large effects control this trait. de Almeida Filho et al. (2016), using the simulated population, reported a slightly lower predictive ability in the oligogenic trait using Bayes RR than Bayes B. In this study, Bayes B and Bayes RR were tested to compare their performance in GWFP because prior distributions and assumptions for both methods are contrasting (Pérez and de Los Campos 2014). Our results showed that both Bayesian methodologies were very similar in predicting family-pools, even for rust incidence in the real population and for the oligogenic trait in the simulated population.

Both cross-validation schemes, leave-one-out and 10-fold, produced similar results in predicting GWFP with a slight advantage for the 10-fold scheme, due to the large variation in the leave-one-out scheme. Resende et al. (2012) reported similar results with the real data set for GEBV, wherein 10-fold and leave-one-out resulted in no significant differences in their predictive ability. Also, similar predictive abilities between the 10-fold and leave-one-out scheme have been reported in wheat (*Triticum aestivum* L.) (Edwards et al. 2019).

Predictive ability of GWFP using contrasting phenotypes

When the families in the validation set had phenotypic values outside the range of phenotypes presented in the training set (bottom and top classes), lower and much more variable predictive abilities were obtained. Interestingly, higher predictive abilities were obtained when families in the validation set had the same phenotypic range as the training set. The impact of the phenotypic variance on prediction was demonstrated by Edwards et al. (2019), which reported that the accuracy of genomic prediction in wheat showed higher predictions for crosses (validation set) with higher phenotypic variance. Würschum et al. (2017) reported equivalent results in triticale (x *Triticosecal* Wittmack), in which higher accuracy was detected for the traits of plant height and biomass in cases in which families with a large phenotypic variation were included in the training/validation set population.

The differences in predictive ability among the scenarios for phenotypic values in the validation set could also be related to the composition of the training sets. For the extreme scenarios (Low and High), the training sets did not have the extreme phenotypic values and alleles frequencies, which could have

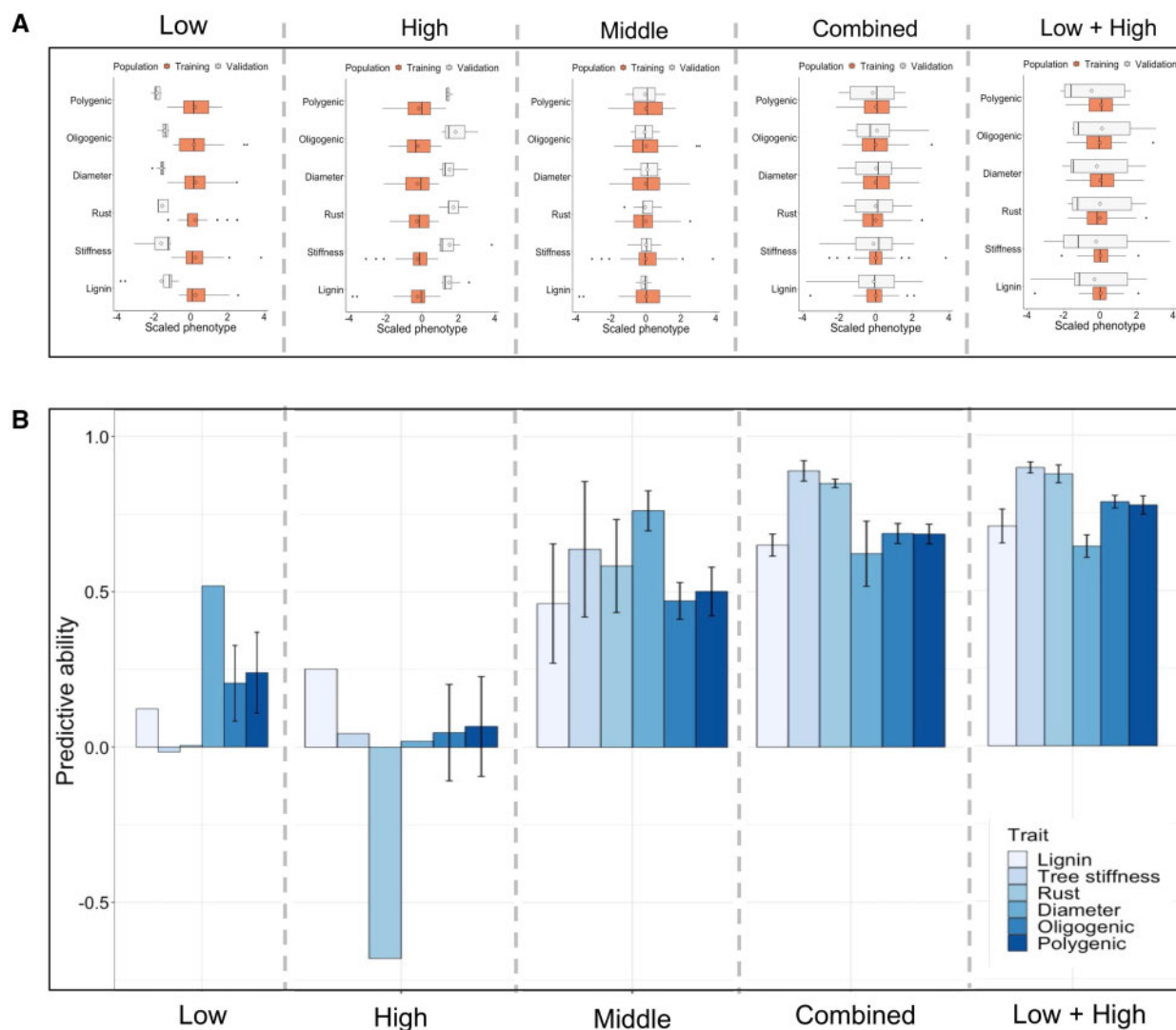


Figure 3 Phenotypic distribution for testing (orange) and validation (white) sets for four traits measured the CCLONES_real population and two traits simulated using CCLONES_sim (A). Average predictive ability obtained with Bayes B using GWFP for four traits in the CCLONES_real (lignin, stiffness, rust, and diameter), and two traits with different genetic architecture (Oligogenic and Polygenic) in the CCLONES_sim populations (B). Five scenarios were tested by creating training (56 families) and validation (7 families) populations using phenotypic data: (i) Low: validation set is composed of seven families with lowest phenotypic records; (ii) High: validation set is composed of seven families with highest phenotypic records; (iii) Middle: validation set is composed of seven families with phenotypic records similar to the family mean; (iv) Combined: two families from Low, two families from High, and three families from Middle; and (v) Low + High: four families from Low and three families from High.

Table 1 Scenarios implemented to design training and validation sets to test predictive ability of genomic prediction models

| Scenario | Set | |
|---------------|-----------------|--|
| | Training | Validation |
| GEBV | 830 individuals | 93 individuals |
| GWFP | 56 families | 7 families |
| GWFP_Fam_Ind | 59 families | 422 individuals |
| GWFP_Fam_Fam | 59 families | 59 families |
| GWFP_Low | 56 families | 7 families with lowest phenotypic values |
| GWFP_High | 56 families | 7 families with highest phenotypic values |
| GWFP_Low_High | 56 families | 7 families, 4 low and 3 high phenotypic values |
| GWFP_Middle | 56 families | 7 families with values similar to the overall mean |
| GWFP_Combined | 56 families | 7 families (2 low, 2 high and 3 middle scenarios) |

GEBV, genomic estimated breeding value; GWFP, genome-wide family prediction; CV, cross-validation.

resulted in poor estimations of markers effects. Studying the optimization process for genomic prediction in wheat, Norman et al. (2018) showed that the genomic prediction accuracy could

be improved, in cases when training set and validation set are not related, by increasing the genetic diversity in the training set.

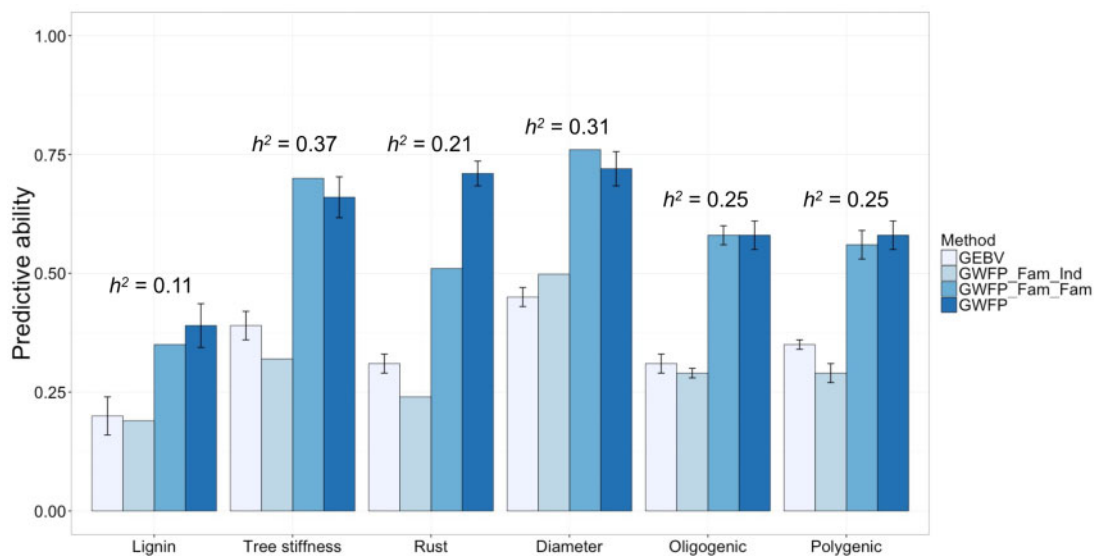


Figure 4 Average predictive ability obtained with Bayes B for four traits in CCLONES-real (lignin, tree stiffness, rust and stem diameter), and two traits with different genetic architecture (Oligogenic and Polygenic) in the CCLONES_sim populations using different genomic prediction methods. GEVB: genomic estimated breeding values individual trees; GWFP_Fam_Ind: genome-wide family prediction using 59 family pools as training set, while different individuals from the same families were used as validation set; GWFP_Fam_Fam: genome-wide family prediction using 59 family pools as the training and validation population, but different full-sib individuals were pooled in both sets; GWFP: genome-wide family prediction using 63 family pools in a 10-fold cross validation scheme. Narrow-sense heritability (h^2) estimated at the individual level (Resende et al. 2012).

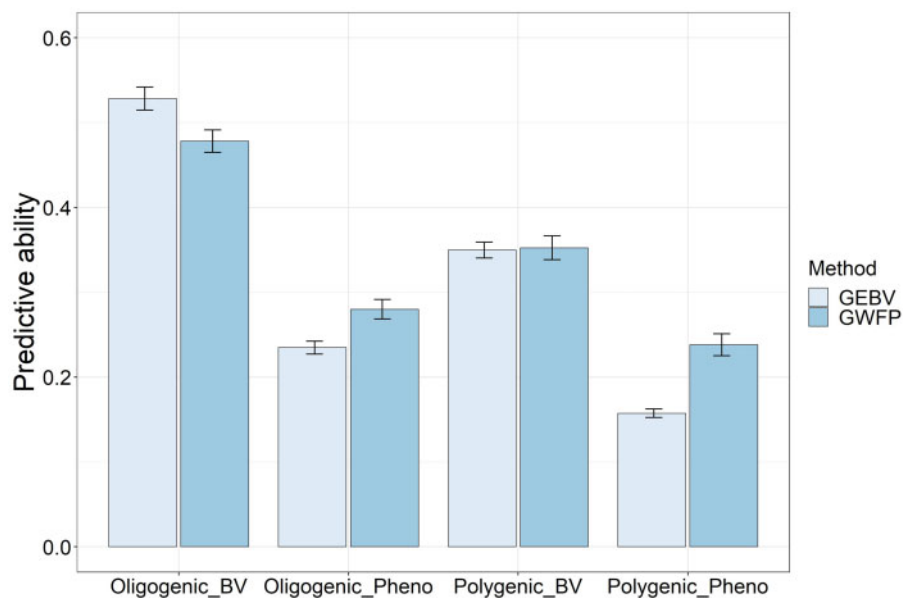


Figure 5 Average predictive ability and accuracy obtained with Bayes B for two traits with different genetic architecture (Oligogenic and Polygenic) in the CCLONES_sim_progeny population, obtained with individual (GEVB) and family-pooled (GWFP) genomic prediction methods. Predictive ability calculated as the correlation between estimated breeding and phenotypic values are denoted as _Pheno, and accuracy as the correlation between estimated and true breeding values as _BV.

Predictive ability of GEVB and GWFP for different traits and scenarios

Predictive ability was always greater for GWFP methods than GEVB in both the real and simulated populations and for all traits, except when the model was built with family pools, and individual performance was predicted (GWFP_Fam_Ind) (Figure 4). Although the full sib families average explores only half of additive genetic variance, the error variance is mitigated with larger number of observations due progeny replication, when compared with single observations (Hallauer et al. 2010). Then, this higher

precision of phenotypic value in family bulks could explain the higher accuracy in genomic prediction of families.

The higher accuracy in the GWFP method was expected since the additive genetic variance explored in this method is just 50% of the additive genetic variance compared with the GEVB. The genotypic value of a family is equal to the mean breeding value of the two parents: $\frac{1}{4}(V_a + V_a) = \frac{1}{2}V_a$ (ignoring the dominance and epistasis effects), so the additive variance among full-sib families is only 50% of the total additive variance, whereas the other 50% represents the variance within a family, which leads to higher accuracy and heritability (Casler and Brummer 2008; Ashraf et al.

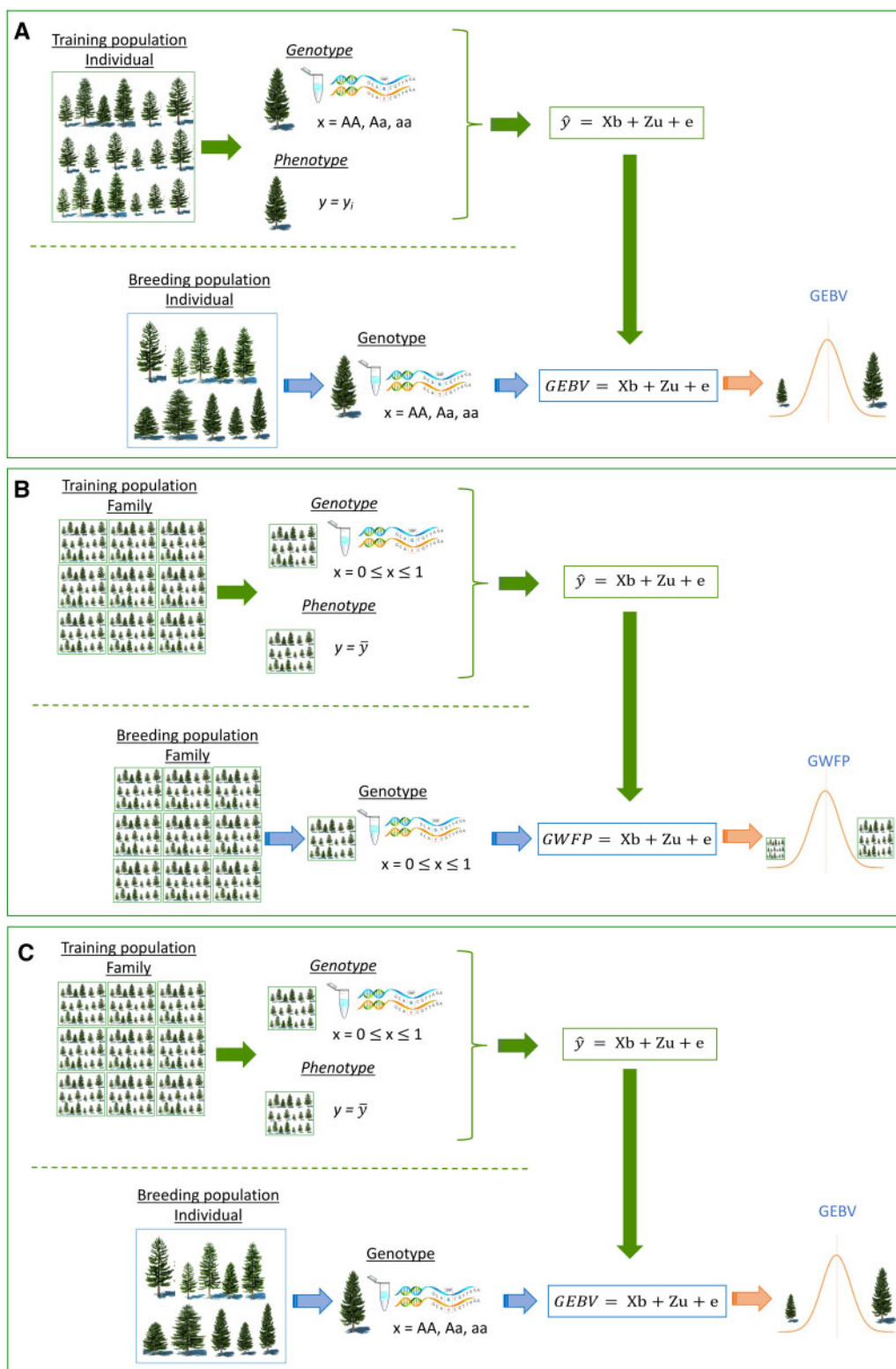


Figure 6 Scheme for the different genomic prediction scenarios: (A) GEBV: genomic estimated breeding values for individual trees; (B) GWFP_Fam_Fam: genome-wide family prediction for families prediction; (C) GWFP_Fam_Ind: genome-wide family prediction applied in the selection of individuals.

2014). Besides, relatedness between the training set and the validation set also influence the predictive ability. The relationship between the training set and the validation set has a crucial role in the model predictive ability (Lorenz and Smith 2015; de Bem Oliveira et al. 2020), it can help explain the higher predictive

ability found in the GWFP_Fam_Fam and GWFP, compared with the GEBV and GWFP_Fam_Ind.

Nevertheless, the predictive ability for most traits obtained with GWFP_Fam_Ind scheme was of the same order of magnitude compared with GEBV, except for the traits stiffness and rust.

Therefore, using the numbers from this study as example, considering the significant reduction in costs incurred in DNA extraction and genotyping 56 families (training set for GWFP), instead of 844 individuals (training set for GEBV), the approach GWFP_Fam_Ind could still be an affordable option for implementing genomic prediction in breeding programs that select individual plants, but have limited budgets to phenotype and genotype all individuals in the training set.

Reduced investments to implementation of genomic prediction with higher predictive ability accuracies can be obtained with the GWFP approach compared with GEBV. A larger number of families can be included in the models, which, for the present population, would likely result in higher predictive abilities as reported in perennial ryegrass for heading date (Fé et al. 2015). Additionally, including more than 10 individuals per family will reduce the sampling variability of the allele frequency and phenotypic mean, resulting in higher genomic accuracies (de Bem Oliveira et al. 2020).

Application of GWFP in a breeding program

Genomic prediction has the power to shorten the time of a breeding process, which leads to a higher genetic gain per unit time, and can allow a reduction in phenotyping process and costs (Grattapaglia and Resende 2011; Crossa et al. 2017; Voss-Fels et al. 2019). However, in some cases, breeders need to genotype a large number of individuals (>10,000) to implement genomic prediction in their programs, increasing costs significantly (Voss-Fels et al. 2019). The high genotyping costs due to large population sizes can make it impracticable to implement genomic prediction in minor crops, particularly in public breeding programs.

For breeding programs with limited budgets, the GWFP can be an alternative to GEBV due to the reduction in phenotypic and genotypic costs to develop prediction models. GWFP has been used in several forage species that are bred in family bulks and whose phenotyping for critical traits is conducted at the sward/plot level (Fé et al. 2015, 2016; Annicchiarico et al. 2015; Biazzi et al. 2017; Jia et al. 2018; Cericola et al. 2018; Guo et al. 2018). In a GEBV approach, the data (phenotypic and genotypic) is collected at the individual level and models are built to estimate the performance of individuals (Figure 6A; Resende et al. 2012; de Almeida Filho et al. 2016, 2019). The GEBV requires significant more resources (labor, economic, and computational) to collect and analyze data. Under a GWFP approach, the number of genotypic samples (bulk DNA and a single-sequencing effort per family) will be the exact number of families, representing a significant reduction in the number of samples compared with the traditional GEBV process (Figure 6B). The phenotyping process will also be performed at the family/plot level, which is the ideal scenario for critical traits in some crops such as forage and turfgrass species.

Breeders may also be interested in employing the GWFP_Fam_Ind approach, where family bulks are used as training set, but individuals are the selection unit (Figure 6C). In this study, the GWFP_Fam_Ind approach showed similar accuracy to GEBV for most traits, with the addition of lower needs for phenotypic and genotypic data for the model development. Finally, GWFP models could be exploited in scenarios when remnant seeds might be available for the same family, and the goal would be to predict the performance of the family or individuals within the family. The remaining seeds from the selected families can be used later to test their merits in further replicated field trials. For perennial allogamous crops, families used in the training set can be used as a new crossing block to start a new selection cycle.

Conclusion

Despite the limitation in number of families and number of individuals per family tested in this study, less than six individuals per family produced inaccurate estimates of family phenotypic performance and allele frequency. Validation sets with similar phenotypic mean and variance as the training set showed greater predictive ability and more accurate predictions consistently across traits. These results revealed great potential for using GWFP in breeding programs that select family bulks as the selection unit, GWFP is well suited for crops that are routinely genotyped and phenotyped at the plot-level. The GWFP approach can also be extended to breeding schemes where family bulks can serve as training sets, while individuals are the selection target.

Data availability

All phenotypic and genotypic data utilized in this study have been previously published as a standard data set for development of genomic prediction methods (Resende et al. 2012; de Almeida Filho et al. 2016). Simulated data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.3126v>.

Conflicts of interest

None declared.

Literature cited

- Amadeu RR, Ferrão Lfv, Oliveira IDB, Benevenuto J, Endelman JB, et al. 2020. Impact of dominance effects on autotetraploid genomic prediction. *Crop Sci.* 60:656–665.
- Annicchiarico P, Nazzicari N, Li X, Wei Y, Pecetti L, et al. 2015. Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics.* 16:1020.
- Ashraf BH, Jensen J, Asp T, Janss LL. 2014. Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. *Theor Appl Genet.* 127:1331–1341.
- Baltunis BS, Huber DA, White TL, Goldfarb B, Stelzer HE. 2007. Genetic gain from selection for rooting ability and early growth in vegetatively propagated clones of loblolly pine. *Tree Genet Genomes.* 3:227–238.
- Barbosa MHP, Resende MDV, Dias LADS, Barbosa GVDS, Oliveira RAD, et al. 2012. Genetic improvement of sugar cane for bioenergy: the Brazilian experience in network research with RIDESA. *Crop Breed Appl Biotechnol.* 12:87–98.
- Biazzi E, Nazzicari N, Pecetti L, Brummer EC, Palmonari A, et al. 2017. Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. *PLoS One.* 12:e0169234.
- Brascamp EW, Bijma P. 2019. A note on genetic parameters and accuracy of estimated breeding values in honey bees. *Genet Sel Evol.* 51:1–6.
- Casler MD, Brummer EC. 2008. Theoretical expected genetic gains for among-and-within-family selection methods in perennial forage crops. *Crop Sci.* 48:890–902.
- Cericola F, Lenk I, Fè D, Byrne S, Jensen CS, et al. 2018. Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* L.). *Front Plant Sci.* 9:369.
- Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res.* 19:136–142.

- 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, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, et al. 2017. Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci*. 22:961–975.
- Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*. 185:1021–1031.
- de Almeida Filho JE, Guimarães JFR, Silva FFE, de Resende MDV, Muñoz P, et al. 2016. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity (Edinb)*. 117:33–41.
- de Almeida Filho JE, Guimarães JFR, Silva FFE, de Resende MDV, Muñoz P, et al. 2019. genomic prediction of additive and non-additive effects using genetic markers and pedigrees. *G3 (Bethesda)*. 9:2739–2748.
- de Bem Oliveira I, Amadeu RR, Ferrão LFV, Muñoz PR. 2020. Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity (Edinb)*. 125:437–448.
- Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, et al. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*. 185:969–982.
- Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, et al. 2019. The effects of training population design on genomic prediction accuracy in wheat. *Theor Appl Genet*. 132:1943–1952.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 6:e19379.
- Esfandiyari H, Fè D, Tessema BB, Janss L, Jensen J. 2020. Effects of different strategies for exploiting genomic selection in perennial ryegrass breeding programs. *G3 (Bethesda)*. 10:3783–3795.
- Falconer DS, Mackay FC. 1996. Introduction to quantitative genetics. In: *Introduction to Quantitative Genetics*. New York: John Wiley & Sons.
- Fè D, Cericola F, Byrne S, Lenk I, Ashraf BH, et al. 2015. Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics*. 16:921.
- Fè D, Ashraf BH, Pedersen MG, Janss L, Byrne S, et al. 2016. Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. *Plant Genome*. 9:1–12.
- Grattapaglia D, Resende MD. 2011. Genomic selection in forest tree breeding. *Tree Genet Genomes*. 7:241–255.
- Gezan SA, Osorio LF, Verma S, Whitaker VM. 2017. An experimental validation of genomic selection in octoploid strawberry. *Hortic Res*. 4:1–9.
- Gianola D. 2013. Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*. 194:573–596.
- Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. 2009. Additive genetic variability and the Bayesian alphabet. *Genetics*. 183:347–363.
- Guo X, Cericola F, Fè D, Pedersen MG, Lenk I, et al. 2018. Genomic prediction in tetraploid ryegrass using allele frequencies based on genotyping by sequencing. *Front Plant Sci*. 9:1165.
- 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*. 42:5.
- Hallauer AR, Carena MJ, Miranda Filho JB. 2010. *Quantitative Genetics in Maize Breeding*. Springer, New York, USA: Springer Science & Business Media.
- Hayes BJ, Daetwyler HD, Bowman P, Moser G, Tier B, et al. 2009. Accuracy of genomic selection: comparing theory and results. *Proc Assoc Advmt Anim Breed Genet*. 18:34–37.
- Hickey JM, Gorjanc G. 2012. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. *G3 (Bethesda)*. 2:425–427.
- Hough J, Williamson RJ, Wright SI. 2013. Patterns of selection in plant genomes. *Annu Rev Ecol Evol Syst*. 44:31–49.
- Jia C, Zhao F, Wang X, Han J, Zhao H, et al. 2018. Genomic prediction for 25 agronomic and quality traits in alfalfa (*Medicago sativa*). *Front Plant Sci*. 9:1220.
- Johnson R, Clair BS, Lipow S. 2001. Genetic conservation in applied tree breeding programs. In: Bart A, Thielges BA, Sastrapradja SD, Rimbawanto A (eds) *Proceedings of the ITTO conference on in situ and ex situ conservation of commercial tropical trees*. ITTO, Yokohama, Japan, pp. 215–230.
- Kumar S, Chagné D, Bink MC, Volz RK, Whitworth C, et al. 2012. Genomic selection for fruit quality traits in apple (*Malus domestica* Borkh. *PLoS One*. 7:e36674.
- Lara LAdC, Santos MF, Jank L, Chiari L, Vilela MDM, et al. 2019. Genomic selection with allele dosage in panicum maximum Jacq. *G3 (Bethesda)*. 9:2463–2475.
- Lin Z, Hayes BJ, Daetwyler HD. 2014. Genomic selection in crops, trees and forages: a review. *Crop Pasture Sci*. 65:1177–1191.
- Lorenz AJ, Smith KP. 2015. Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Sci*. 55:2657–2667.
- Massman JM, Jung HJG, Bernardo R. 2013. Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. *Crop Sci*. 53:58–66.
- Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157:1819–1829.
- Munoz PR, Resende MFR, Huber DA, Quesada T, Resende MDV, et al. 2014. Genomic relationship matrix for correcting pedigree errors in breeding populations: impact on genetic parameters and genomic selection accuracy. *Crop Sci*. 54:1115–1123.
- Norman A, Taylor J, Edwards J, Kuchel H. 2018. Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3 (Bethesda)*. 8:2889–2899.
- Pembleton LW, Drayton MC, Bain M, Baillie RC, Inch C, et al. 2016. Targeted genotyping-by-sequencing permits cost-effective identification and discrimination of pasture grass species and cultivars. *Theor Appl Genet*. 129:991–1005.
- Pembleton LW, Inch C, Baillie RC, Drayton MC, Thakur P, et al. 2018. Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass. *Theor Appl Genet*. 131:1891–1902.
- Pérez P, de Los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. 198:483–495.
- Pérez-Cabal M, Vazquez AI, Gianola D, Rosa GJ, Weigel KA. 2012. Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Front Genet*. 3:27.
- Poehlman JM. 1987. *Breeding cross-pollinated and clonally propagated crops*. In: *Breeding Field Crops*. Dordrecht: Springer, p. 214–236.
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu SY, et al. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*. 5:103–113.

- R Core Team, 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. URL <http://www.R-project.org/>.
- Resende MDVD, Barbosa MHP. 2006. Selection via simulated individual BLUP based on family genotypic effects in sugarcane. *Pesq Agropec Bras.* 41:421–429.
- Resende MF, Muñoz P, Resende MD, Garrick DJ, Fernando RL, *et al.* 2012. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics.* 190:1503–1510.
- Stock KF, Reents R. 2013. Genomic selection: status in different species and challenges for breeding. *Reprod Dom Anim.* 48:2–10.
- Torres LG, Vilela de Resende MD, Azevedo CF, Fonseca e Silva F, de Oliveira EJ. 2019. Genomic selection for productive traits in biparental cassava breeding populations. *PLoS One.* 14:e0220245.
- VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, *et al.* 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci.* 92:16–24.
- Vencovsky R, Crossa J. 2003. Measurements of representativeness used in genetic resources conservation and plant breeding. *Crop Sci.* 43:1912–1921.
- Voss-Fels KP, Cooper M, Hayes BJ. 2019. Accelerating crop genetic gains with genomic selection. *Theor Appl Genet.* 132:669–686.
- Wang Q, Yu Y, Yuan J, Zhang X, Huang H, *et al.* 2017. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp *Litopenaeus vannamei*. *BMC Genet.* 18:1–9.
- Wang J, Cogan NO, Forster JW. 2016. Prospects for applications of genomic tools in registration testing and seed certification of rye-grass varieties. *Plant Breed.* 135:405–412.
- Wetterstrand KA, 2020. DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). National Human Genome Research Institute, National Institutes of Health, Bethesda, Md. <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data> (Accessed: 2021July 19).
- Würschum T, Maurer HP, Weissmann S, Hahn V, Leiser WL. 2017. Accuracy of within-and among-family genomic prediction in triticales. *Plant Breeding.* 136:230–236.
- Xu Y, Liu X, Fu J, Wang H, Wang J, *et al.* 2020. Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Commun.* 1:100005.

Communicating editor: A. E. Lipka