Euphytica (2024) 220:146
https://doi.org/10.1007/s10681-024-03406-2

RESEARCH

Comparing strategies for genomic predictions in interspecific biparental populations: a case study with the *Rubus* genus

Allison Vieira da Silva · Melina Prado · Gabriela Romêro Campos · Karina Lima Reis Borges · Rafael Massahiro Yassue · Gustavo Husein · Marcel Bellato Sposito · Lilian Amorim · José Crossa · Roberto Fritsche-Neto

Received: 14 April 2024 / Accepted: 27 August 2024

Abstract Genomic selection (GS) is becoming increasingly widespread and applied due to the promising results obtained, cost savings in generating single nucleotide polymorphism (SNP) markers, and the development of statistical models that allow to improve the analysis robustness and accuracy. The composition and size of the training population have a major influence on GS, which poses challenges for interspecific biparental populations. Another factor is the use of different reference genomes from other species to perform SNP calling, which could make it possible to explore variability in interspecific crosses comprehensively. Late leaf rust is a disease caused by the pathogen *Acculeastrum americanum,* and there are reports on genetic resistance in *Rubus occidentalis,* which leads to the need for interspecific hybridizations, aiming to combine the fruit quality of *R. idaeus* with the resistance of *R. occidentalis*. The present study was carried out with a population of 94 interspecific raspberry hybrids. We evaluated the effect of different reference genomes on the SNP markers discovery, as well as training population optimization strategies on the accuracy of genomic predictions, namely the CV-α, leaving-one-family-out (LOFO), pairwise families, and stratified k-fold. The average predictive accuracies ranged from −0.33 to 0.44 and We demonstrated higher prediction accuracy and more precise estimates when we combined stratified sampling to compose the training set (CV-α and k-fold stratified CV) and the panel of Unique markers. These results corroborate that genomic prediction aligned with SNP calling and training population optimization strategies can significantly increase genetic gains in interspecific biparental crosses.

Keywords Interspecific · Hybrids · Training set · Genomic selection · *Rubus*

Abbreviations

ANOVA	Analysis of variance
CV	Cross-validation
GBLUP	Genomic best linear unbiased predictor

A. V. da Silva (✉) · M. Prado · G. R. Campos · G. Husein
Department of Genetics, Luiz de Queiroz" College of Agriculture, University of São Paulo (USP), Piracicaba, São Paulo, Brazil
e-mail: allisonvsagro@usp.br

K. L. R. Borges · J. Crossa · R. Fritsche-Neto
Rice Research Station, Louisiana State University AgCenter, Rayne, USA

R. M. Yassue
GDM Seeds, Campinas, São Paulo, Brazil

M. B. Sposito
Crop Science Department, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, SP 13418-900, Brazil

L. Amorim
Department of Plant Pathology and Nematology, Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP), Piracicaba, São Paulo, Brazil

Published online: 04 September 2024

GBS	Genotyping by sequencing
GS	Genomic selection
LD	Linkage disequilibrium
LOFO	Leave-one-family-out
PF	Pairwise family
SNP	Single nucleotide polymorphism
SS	Sum of squares
TS	Training population

Introduction

Every year, genomic selection (GS) (Bernardo 1994; Meuwissen et al. 2001) becomes more widespread and applied due to the promising results obtained, cost reduction in the generation of SNP markers, and the development of statistical models that allow the inclusion of more and more data (Crossa et al. 2017; Lebedev et al. 2020; Montesinos-López et al. 2023). In addition, GS makes it possible to shorten the selection cycle, which has a major impact, especially for perennial species, allowing early selection of plants that are still small and before fruiting (Kainer et al. 2015; Iwata et al. 2016; Lebedev et al. 2020). The application of GS in perennial species saves physical space and maintenance costs in trials since superior genotypes are early selected and all efforts are focused on the selected individuals (Kainer et al. 2015; Fritsche-Neto et al. 2012).

For genomic prediction, the models are developed from a set of genotyped and phenotyped individuals, which form the training population (TS) and applied to the test population, containing individuals that connect the two populations by kinship, making it possible to obtain an estimate of the breeding values of the individuals in the breeding population without the need to know the phenotype of these individuals, using only genotypic data and the genetic relationship between individuals (Desta and Ortiz 2014; Kwong et al. 2017; Xu et al. 2019). Selection based solely on genotypic data can be carried out in the early developmental stages, speeding up the breeding program by shortening the crop cycle (Xu et al. 2019; Montesinos-López et al. 2023).

The implementation of GS in biparental interspecific populations becomes complex (Olatoye et al. 2020). In this scenario, the composition and size of the training population have the greatest influence on predictive ability, as genetic structure within families and between species are common (Tan et al. 2017). In addition, interspecific hybridization is a naturally occurring phenomenon that generates great genetic variability, new haplotypes, and major changes in population allele frequencies and is one of the main mechanisms responsible for plant adaptation to abrupt environmental changes and the emergence of new species (Runemark et al. 2019). These factors must be considered when building the training population to mitigate bias in predictive models. Another possibility arising from using these interspecific populations is the availability of alternative reference genomes from the different species to perform SNP calling, which may allow the variability and compatibility of crosses to be explored (Lara et al. 2019).

In this context, the best-known raspberry species are the black raspberry (*Rubus occidentalis*) and the red raspberry (*Rubus idaeus*), with the red ones being economically more important and more consumed than the black ones (Baby et al. 2018). In addition, the most widely cultivated varieties worldwide have a narrow genetic base. Wild relatives occur in diverse environments, and this variability represents a genetic resource available for study and the development of new cultivars (Hall et al. 2009). Specifically in this crop, late leaf rust is a disease caused by the pathogen *Acculeastrum americanum*, which is responsible for causing premature defoliation, increasing susceptibility to winter damage. The pathogen also infects the fruit, making it unsuitable for the fresh fruit market. The disease begins with small orange spots on the abaxial part of the leaf, turning brown over time. Young leaves are the last to show disease symptoms; middle-aged leaves are the most susceptible (Ellis et al. 1991; Hall et al. 2009). The literature on the genetic basis of disease resistance in the "Jewe"' variety is scarce, as well as adaptations to tropical climatic conditions. The rare reports on genetic resistance to this pathogen are in *Rubus occidentalis* species (Hall et al. 2009), which leads to an urgent need for interspecific hybridization aiming to combine the fruit quality of the *R. idaeus* species with the resistance of the *R. occidentalis* species. Prado et al. (2024) in a recent study have found evidence that resistance to raspberry late leaf rust is polygenic, with regions of major effects and minor effects that play significant roles in rust resistance in raspberry.

Given the above, the objective was to evaluate the effect of the reference genome for the discovery of SNP markers and the training population composition strategies on the accuracy of genomic predictions models for late leaf rust in biparental and interspecific populations of the *Rubus* genus.

Material and methods

Biological material

Ninety-four raspberry hybrids were obtained from crosses in partnership with the Plant Production department at ESALQ/USP. The crosses were made using a *testcross* scheme, in which the first group consisted of three parents of the *Rubus idaeus* species of the "Golden Bliss", "Salmon", and "Himbo Top" varieties, with favorable morpho-agronomic characteristics and different levels of susceptibility to late leaf rust. The second group comprised *Rubus occidentalis* parents of the "Jewel" variety, a source of late leaf rust resistance alleles. Different numbers of hybrids were obtained from each cross, and in all crosses, the "Jewel" variety was used as the female parent (Table 1).

More details about the materials used can be found in the work published by Campos et al. (2023). In this work, the authors characterized this panel of hybrids' diversity and genetic structure, which initially had 116 genotypes. Still, because they are temperate climate materials, some of the genotypes were lost in the process of adapting to the climatic conditions found in Piracicaba, São Paulo, Brazil (22°42′S, 47°38′W, 540 m), where the experiment was conducted. In addition to climatic adversities, genetic incompatibility may have led to the loss of some materials.

Table 1 Number of hybrids (N) obtained from crosses between the Jewel variety, as the female relative in all crosses, and the Golden Bliss (JG), Salmon (JS), and Himbo Top (JT) varieties

Parent	Jewel	Golden Bliss	Salmon	Himbo Top
Family	–	JG	JS	JT
N	–	35	28	31

Inoculation

The inoculum of the fungus *Acculeastrum americanum* was prepared by the Phytopathology Department of ESALQ/USP. Suspensions were obtained using 50 mL of distilled water, Tween 20 (0.01%), and *A. americanum* urediniospores. The suspension concentration was adjusted to 10^4 urediniospores/mL in a Neubauer chamber and used to inoculate by spraying the abaxial side of the leaves up to the point of oozing. In order to guarantee the development of the disease, the pots were covered for 24 h with a plastic bag to create a humid chamber.

Phenotyping and conducting the experiment

The experimental unit was defined as a single plant in a 5-L pot. The plants were grown in a greenhouse in Piracicaba-SP, Brazil (22°42′S, 47°38′W, 540 m). The experimental units were arranged in an augmented block design repeated across time, with three blocks, where the parents ("Golden Bliss," "Salmon," "Himbo Top," and "Jewel") were used as checks and were present in each one of the blocks. The experiment was repeated twice, once in November 2021 and again in April 2022. Phenotyping was carried out based on the classification of the plants in terms of the severity of the disease on the leaves of the plants. The severity of the disease was classified on the seventeenth day after inoculation based on the diagrammatic scale proposed by Dias et al. (2022) with eight levels, from 0 to 8, with 0 being the absence of the disease and 8 being the most severe stage. Three measurements were taken for each plant by three different evaluators, and an arithmetic mean was generated from the three observations to establish the final classification.

Genomic characterization

The DNA from the samples was extracted from the leaf tissue according to the protocol suggested by the manufacturer of the extraction kit (Qiagen), using the DNeasy Plant Mini Kit. After extraction, the genomic library was built based on the protocol proposed by Poland et al. (2012) with some adaptations. The enzymes PstI (rare cut) and MseI (frequent cut) from

New England BioLabs Inc.® were used to digest the DNA. Libraries were sequenced on a HiSeq 2500 System sequencer (Illumina, Inc).

The samples were sequenced on the Illumina platform in collaboration with the Genetic Diversity and Improvement Laboratory of the Genetics Department at ESALQ/USP. SNP calling was carried out using TASSEL-GBS (Glaubitz et al. 2014). The sequences were aligned using two reference genomes available for raspberries: the genome of the red raspberry, *Rubus idaeus* (scaffolds), and the genome of the black raspberry, *Rubus occidentalis* (chromosomes). A total of 275,904,265 sequences were identified and submitted to quality control. Only SNPs with a minimum allele frequency (MAF) of 20% and a call rate of over 90% were selected. This MAF value was used to control possible errors from the genotyping platform and because we were interested in the resistance alleles of the maternal progenitor. Missing data was imputed using the kNNI method from the impute package (Hastie et al. 2022). In the filtering for missing data, no individuals with a percentage of missing data greater than 30% were identified.

After quality control, three marker matrices were generated. The Moc matrix, with 20,382 SNPs, was obtained by aligning the sequences to the reference genome of the *R. occidentalis* species. The Mid matrix, with 20,133 SNPs, was generated by aligning the sequences to the genome of the *R. idaeus* species. The third matrix with 30,398 SNPs, the Unique matrix, was obtained by joining the Moc and Mid matrices but eliminating redundant markers between the two matrices (Campos et al. 2023). The markers were also filtered for the LD (Linkage Disequilibrium) parameter; the LDs between the markers were calculated between 100 Kbp intervals using the correlation method, and the 99% threshold was used to filter the markers. The three marker matrices used were the same as those used in a previous work published by Campos et al. (2023) in which, using principal component analysis, we evaluated how the hybrid population is grouped according to each of the matrices.

Phenotypic analysis

The phenotypic data was analyzed using the statgenSTA package (Rossum et al. 2023) in the R environment using the following model:

$$\mathbf{y} = \mathbf{X_1 r} + \mathbf{X_2 g} + \mathbf{Z b} + \mathbf{e} \quad (1)$$

where **y** is the vector of observed severity values in a given stage; **r** is the fixed effect of repetition; **g** is the fixed effect of genotype; **b** is the random effect of block nested with repetition, where $\mathbf{N}(0, \mathbf{I}\sigma_b^2)$; **e** is the random effect of the residuals, where $\mathbf{N}(0, \mathbf{I}\sigma_e^2)$; X_1 and X_2 are the fixed effect incidence matrices; $\mathbf{Z_1}$ is the random effect incidence matrix; **I** is an identity matrix. The adjusted means for each genotype and stage from this model were used in the subsequent analyses.

Prediction models

Two prediction models were tested, the additive and the additive-dominant (VanRaden 2008):

$$\mathbf{y} * = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\delta} + \mathbf{e} \quad (2)$$

where ***y**** corresponds to the adjusted measurements of the individuals in the first stage, with dimension $n \times 1$ where n corresponds to the number of observations; **X** corresponds to the matrix of incidence of the fixed effects, **β** corresponds to the vector of the fixed effect of the intercept, **Z** corresponds to the matrix of incidence of the random (genetic) effects with dimension n x n, where n is equal to the number of genotypes, **α** and **δ** correspond to the vectors related to **Z** of the effects of dominance and additivity, respectively, where $\boldsymbol{\alpha} \sim N(0, \mathbf{Ga}\ \sigma_a^2)$, $\boldsymbol{\delta} \sim N(0, \mathbf{Gd}\ \sigma_d^2)$, **Ga** and **Gd** are genetic relationship matrices of additivity and dominance, respectively, obtained by the methodology described by VanRaden (2008) using the snpReady package (Granato et al. 2018). The residual is represented by **e,** where $\mathbf{e} \sim N(0, \mathbf{I}\ \sigma_e^2)$. The prediction models were implemented in the R software with the help of the BGLR package (Perez and Los Campos 2014) using 15,000 burn-ins and 30,000 iterations.

Genomic heritability in the restricted sense was calculated based on the variance components obtained with the G matrix from fitting the additive GBLUP (Genomic Best Linear Unbiased Predictor) model in the BGLR package. We used the ratio between the genetic variance based on the markers and the sum of the genetic variance based on the markers and the residual variance. Heritability was estimated for each of the marker panels.

Table 2 Factors considered when evaluating the accuracy of the prediction models tested

Factors		
CV	Panel	Model
CV-α	Mid	A
LOFO	Moc	AD
PF	Unique	
K-fold		

Four cross-validation (CV) strategies were evaluated: the CV-α method (Yassue et al. 2021), leave-one-family-out (LOFO), pairwise families (PF), and stratified k-fold (STRAT). The three marker matrices (Panel) are represented by Mid (generated from the reference genome of the *R. idaeus* species), Moc (generated from the reference genome of the *R. occidentalis* species), and Unique (generated from combining the unique markers present in the other two matrices). A and AD, respectively, represent the additive and additive-dominant models

Cross-validation of prediction models

Four cross-validation schemes were used (Table 2). In the first (Fig. 1a) we used the CV-α method (Yassue et al. 2021) with five folds and four repetitions. In the second method (Fig. 1b), the K-fold was stratified with five folds, in which each family contributed 20% of its individuals to the composition of the training population. In the third, LOFO (leave-one-family-out), one family was used as the validation population, and the other two families constituted the training population, with a total of three repetitions (Fig. 1c). In the fourth scheme (Fig. 1d), (PF) we separated the families into pairs and used one of the families as the TS and the other as the validation population in all six possible combinations between the three families (JG/JS, JS/JG, JG/JT/, JT/JG, JS/JT and JT/JS). In each scenario, the accuracy of the prediction was

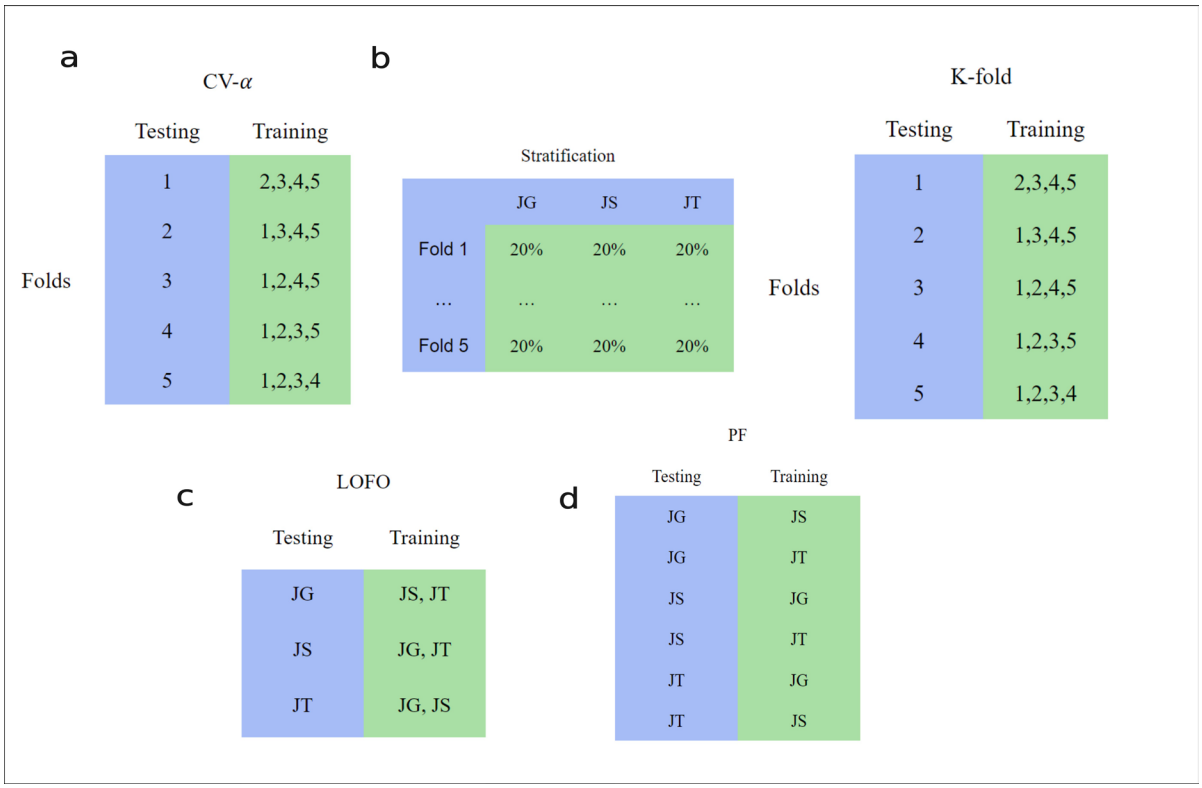

Fig. 1 Cross-validation schemes evaluated in the study. The number (1–5) indicates the folds and JG (Jewel x Golden Bliss), JS (Jewel x Salmon), JT (Jewel x Himbo Top) the crossings. **a** CV-α, **b** K-fold with stratification balanced by family keeping 20% and 80% of the families in the testing and training set, respectively, **c** LOFO (leave-one-family-out), **d** PF (pairwise family)

estimated using Pearson' correlation between the adjusted mean values and the values estimated by the GBLUP model.

An analysis of variance was carried out to assess the influence of each factor on the mean prediction accuracies. The effect size corresponds to the eta-squared calculated for each of the effects from the percentage corresponding to the sum of the squares (SS) of each effect concerning the total sum of the squares according to the following equation:

$$\text{Eta squared} = \frac{\text{SS effect}}{\text{SS total}} \quad (3)$$

Results

The hybrids showed great phenotypic variability in terms of disease severity. The values ranged from the absence of the disease to the maximum level of severity among the genotypes, and variation could be observed within all the families (Fig. 2). The calculated heritabilities for the Mid, Moc, and Unique matrices were 0.37, 0.38, and 0.33, respectively.

The average predictive accuracies ranged from −0.33 to 0.44. The results obtained using the additive and additive-dominant models (described by tables A and AD in Fig. 3) show a similar pattern in terms of mean and data dispersion when combined with the panel of brands and the cross-validation method.

The different marker panels showed variation in terms of mean and distribution depending on the cross-validation method used. In the LOFO and PF cross-validation methods, the three marker matrices also showed a wide dispersion in prediction accuracy across folds, although PF generated a higher prediction accuracy mean (Fig. 3). Although in the ANOVA, only the SNP panel appeared to be significant in the average accuracy values obtained (Table 3), we could see that the Unique matrix provides better accuracy when combined with CV methods which do not consider family structure. In the CV-α method, we observed the smallest dispersion of results around the mean. Using the CV-α and K-fold methods, the Unique matrix results indicated a gain in prediction accuracy compared to the other two marker matrices.

Discussion

Ideally, a prediction model should enable GS on individuals from several different populations while maintaining satisfactory accuracy values for the target trait. One of the assumptions of the GBLUP model, and the majority of regression methods, is that the effect of allelic substitution is homogeneous for all traits among all individuals (de los Campos & Sorensen 2014). However, prediction accuracy is influenced by several factors, such as the size and genetic diversity of the training population, the heritability of the trait of interest, the density of the marker panel, the effects of markers and genes on the trait, the extent and distribution of LD between markers and QTLs (Kaler et al. 2022), the presence of population structure, differences in the linkage phases of haplotypes and large variations in allele frequencies in the different subpopulations or families in the training set composition of the population to be predicted.

In a preliminary study, Campos et al. (2023) observed that the three marker panels performed similarly in assessing genetic diversity and structure parameters. In our study, we observed different performances regarding prediction accuracy between the marker panels, where the Unique matrix performed best. One aspect that should be taken into account is the different sizes of the marker panels: the Unique matrix has 30,398 SNPs, the Moc matrix has 20,382 SNPs, and the Mid matrix has 20,133 SNPs. A greater density of markers can improve genomic predictions but also results in higher genotyping costs, so it is important to balance the gains in prediction accuracy with the costs associated with obtaining a denser panel (DoVale et al. 2022). In addition, it is commonly agreed that the gains in prediction accuracy from increasing the number of markers eventually reach a plateau (Krishnappa et al. 2021). Using different reference genomes from different species to perform SNP calling has made it possible to explore variability in interspecific crosses comprehensively (Lara et al. 2019). In this context, the use of a marker panel based on the two genomes may have made it possible to capture rare variants of lesser effect and different haplotypes at different linkage stages. Thereby, this marker panel probably better captured the effect of allelic substitution and

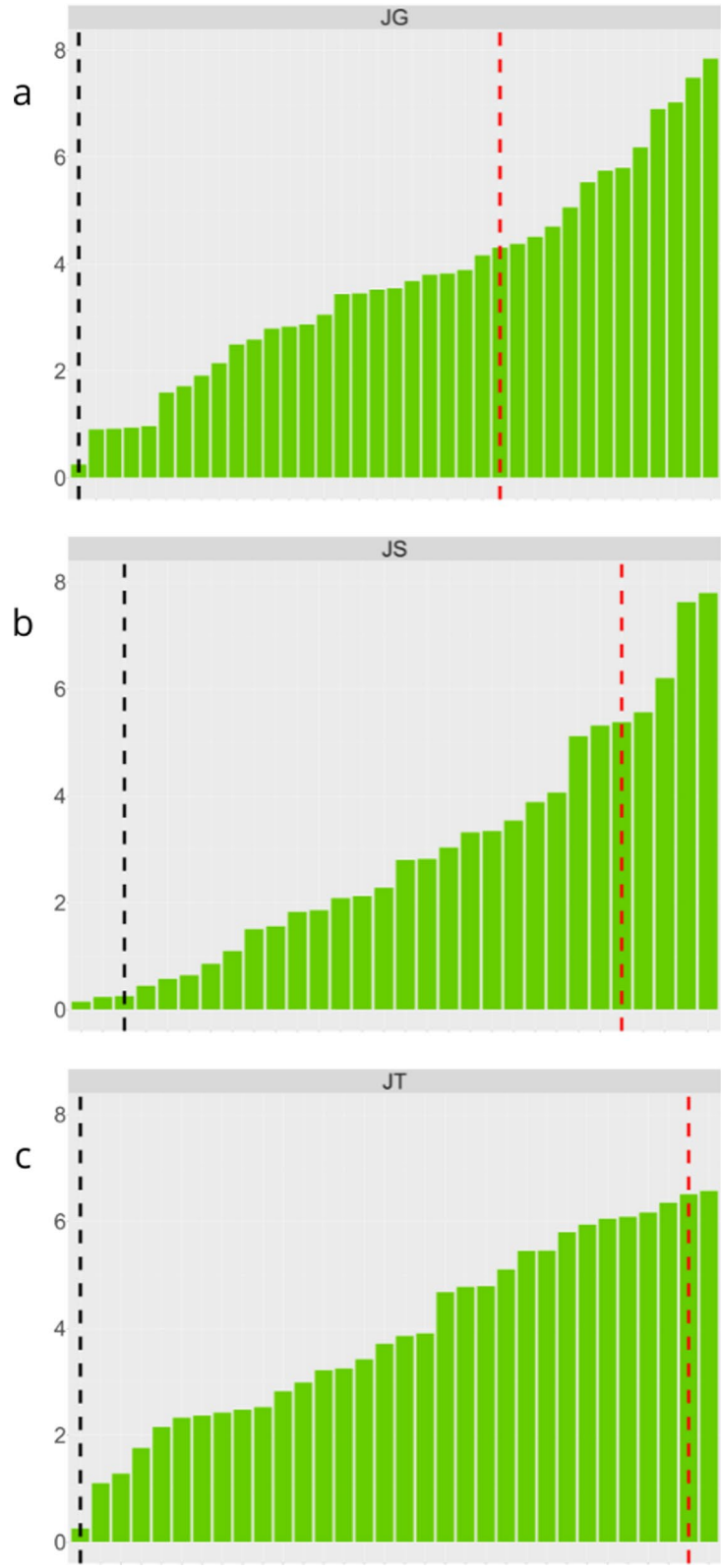

Fig. 2 Barplot in ascending order of the distribution of values for disease severity on the Y axis and each of the hybrids and the parents of each family on the X axis. **a** distribution of severity values for the JG family, where the vertical line in black represents the value presented by the "Jewel" parent (0.26) and the line in red represents the Golden Bliss parent (4.31). **b** distribution of severity values for the JS family, where the vertical black line represents the value presented by the "Jewel" parent (0.26) and the red line represents the Salmon parent (5.38). **c** distribution of severity values for the JT family, where the vertical black line represents the value presented by the "Jewel" parent (0.26) and the red line represents the Himbo Top parent (6.51)

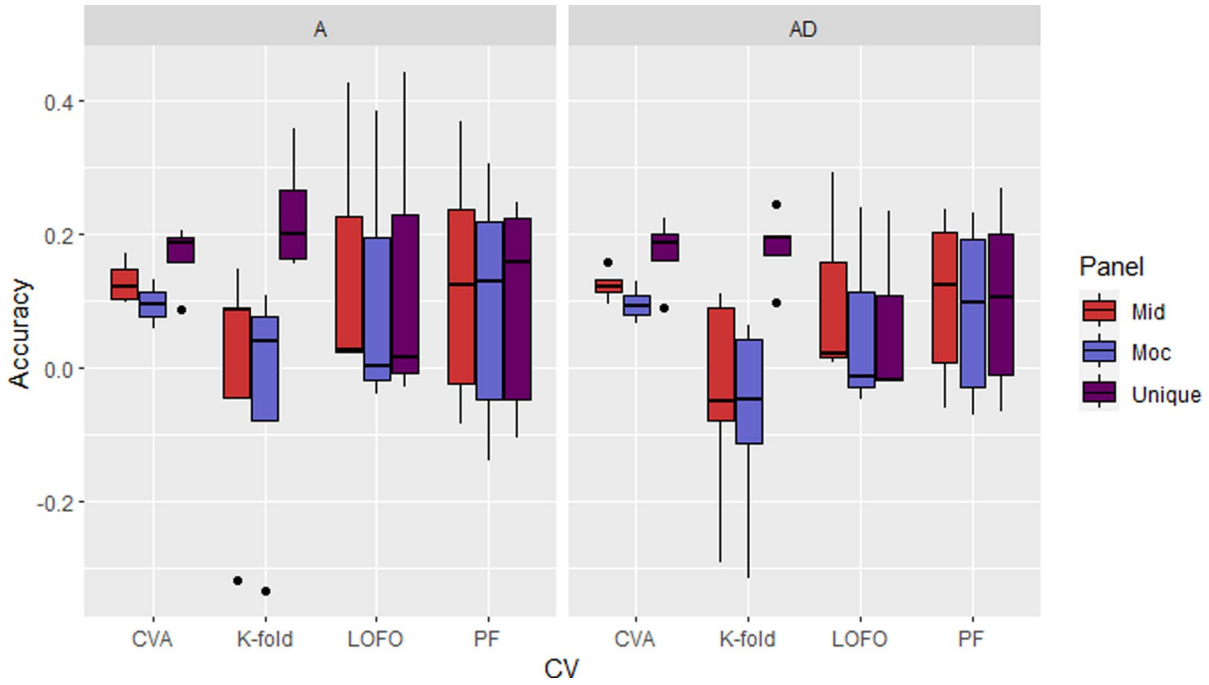

Fig. 3 Accuracy of severity prediction as a function of the CV-A (CV-α), K-fold (with five folds and 20% representation of each of the three families) LOFO (leave-one-family-out), and PF (pairwise families) cross-validation schemes, respectively. The values obtained from the additive model are on the left, represented by A, and the values from the Additive and Dominant model are on the right, represented by AD. The variation in the coloring of the box plots represents the three panels of SNPs

Table 3 Analysis of variance carried out to assess the influence of each factor on the mean prediction accuracies

Source	Df	Sum Sq	F	Pr(> F)	Eta-squared
CV	3	0.1198	2.000	0.1188	0.0511
Panel	2	0.1676	4.194	0.0178*	0.0722
Model	1	0.0155	0.775	0.3808	0.0066
Residuals	101	2.0176			

CV corresponds to the cross-validation scheme, Panel corresponds to marker panel used (Mid, Moc, or Unique), and model corresponds to the prediction model (additive GBLUP or additive and dominant GBLUP). The size effect corresponds to the eta-squared calculated for each effect (Cohen 1988)

*Statistically significant with $p < 0.05$

consequently improved the accuracy (Rooney et al. 2022; MacLeod et al. 2016).

Genomic selection in populations with family structure becomes complex and challenging, as with the interspecific population used (Tan et al. 2017; Olatoye et al. 2020). The presence of genetic structure can reduce the stability and predictive accuracy of a model across populations due to different aspects. Lehermeier et al. (2015) observed differences in marker effect estimates between different clusters in the same population due to differences in the LD pattern of markers and QTL. Legarra et al. (2021) observed that allelic substitution effects could vary between populations and across generations due to changes in genetic relationships, magnitude of additive and non-additive variances and allele frequencies. In this scenario, the composition and size of the training population have a major impact on the predictive capacity of the model since the model tends to perform better if trained on a group of individuals that best represent all these aspects of population diversity (Isidro et al. 2015; Tan et al. 2017; Berro et al. 2019; Montesinos-López et al. 2024). These factors must be considered when building the training population in order to mitigate bias in predictive models.

Raspberry is an allogamous species, and it is important to highlight that in the specific case of this population under study, the F1 segregating

generation was generated from a set of interspecific crosses. Nevertheless, there is a lack of background knowledge on the genetic architecture of resistance against leaf late rust disease in raspberries. In this scenario, we sought to expand the GBLUP model to evaluate the inclusion of the dominance effect in the GBLUP model. The use of this model, considering the effects of additivity and dominance, has had a positive impact on gaining prediction accuracy and selecting elite clonally propagated materials (Resende et al. 2017; Nadeau et al. 2023) such as raspberry. However, we observed that the inclusion of the dominance effect did not generate gains in prediction accuracy compared to the additive model. It may be due to the absence of a significant dominance effect or other influencing factors. The total size of the population and the number of individuals per family are factors that influence the estimation of the dominance effect and, consequently, can impact the gain in prediction accuracy from the inclusion of dominance (Tan et al. 2018). The limited number of individuals in our populations makes it difficult to accurately estimate dominance effects. So, we did not observe any differences between the accuracies obtained with and without the inclusion of the effect.

Regarding the cross-validation scheme and the composition of the training population, we observed higher prediction accuracies when we combined the CV-α and the K-fold method with stratified sampling for the composition of the TS together with using the Unique brand matrix. Although the panel was the only statistically significant factor according to the ANOVA analysis, we observed that the Unique matrix generated similar accuracies to the other matrices when we considered the family structure (LOFO and PF) in the cross-validation scheme. Due to the larger number of markers, we expected the Unique panel to improve prediction accuracy across all CV schemes. However, our results suggest that the increase in the number of markers was not sufficient to compensate for the lower relatedness between the training and testing sets in the LOFO and PF schemes. The CV-α method provided estimates of prediction accuracy with less dispersion than the other CV methods evaluated, as the main purpose of the method is to allow genotypes to be allocated to folds in such a way as to maximize the independence of accuracy estimation errors (Yassue et al. 2021). The K-fold method with a stratified composition of the training population can be a tool used to minimize the effect of genetic structure by sampling in proportion to the size of each cluster, thus potentially capturing genetic diversity in the training population and improving the model's predictive capacity (Isidro et al. 2015; Hoque et al. 2024). The K-fold method considers balanced sampling across families, with 20% of the individuals from each family being sampled to form each of the five folds. The sampling of individuals within each family was done randomly. In contrast, CV-α sampling remains random and does not account for the existence of three different families or their sizes. Our population is small, and we have different numbers of individuals within each family. Although the numbers are relatively close, the JG family has 35 individuals, while the JS family has 28, meaning the JG family has 25% more individuals than the JS family. These differences in family size and sampling methodology may explain the variations in prediction accuracy between the methods with random sampling.

Overall, our results highlight the importance of carefully designing the training set for raspberry breeding when using genomic selection to make predictions across populations. When dealing with structured populations or family structures, it is crucial to balance and stratify the breeding population into training and testing sets to minimize potential bias in effect estimation, especially in small populations. This approach helps reduce the risk of encountering negative accuracies observed in pairwise and leave-one-family-out (LOFO) cross-validations. Strategies such as test-and-shelf may present a viable alternative for implementing genomic selection in raspberry breeding (Boyles et al. 2024). On the other hand, these strategies require much larger and more diverse training sets, and this aspect should be carefully analyzed to balance the costs of genotyping and phenotyping a larger number of individuals against the gains in prediction accuracy (Wu et al. 2023).

Genetic structure can arise from different levels of genetic relatedness between individuals, including separating individuals into families (Würschum et al. 2017). Studies show that genomic prediction within families can generate significant gains in prediction accuracy in the presence of different patterns of LD of the markers, allele frequencies, and different substitution effects

(Würschum et al. 2017; Berro et al. 2019). Thus, making predictions within raspberry families would be a useful tool to elucidate the impact of genetic structure at the family level. However, one of the biggest challenges in making predictions in raspberry is the difficulty in obtaining and maintaining hybrids since the limited number of individuals commonly sampled makes it impossible to build training and validation populations with a satisfactory number of individuals (Montesinos-López et al. 2024). In regions with a hot and humid climate, raspberries have limitations regarding vegetative development, flower production, and fruit set. Long-term exposure to stress caused by high temperatures can inhibit photosynthesis and cause premature plant death (Fernandez et al. 2018). This study was carried out in the municipality of Piracicaba, State of São Paulo, Brazil (22°42′30″ S, 47°38′00" W, 546 m). The location has a tropical climate with a dry winter season, classified as *Aw* in the Köppen climate classification (Dias et al. 2017). The experiment was planned to accommodate more individuals, with more than 160 hybrids obtained from the crosses. Still, unfortunately, many of the individuals did not adapt to the tropical climatic conditions and were lost during the planning phase of the work. In addition to the difficulty in obtaining viable individuals from interspecific crosses due to gametic incompatibility (Pinczinger et al. 2021).

With this work, we have provided the community with an initial attempt to implement GS in a population of interspecific raspberry hybrids. Among the factors that added complexity to the development of the study, we can highlight the limited size of the populations analyzed and the scarce literature on the genetic basis of disease resistance in the "Jewel" variety, as well as adaptations to tropical climatic conditions. We demonstrated higher prediction accuracy and more precise estimates when we combined stratified sampling to compose the training set (CV-α and k-fold stratified CV) and the panel of Unique markers. We have provided important information on the complexity of efficiently sampling the genetic diversity of the genomes of the two species and a first direction in developing a strategy for constructing TS. Additionally, we demonstrated the effect of population structure under different CV schemes in interspecific raspberry hybrids.

Author contributions AVS, JC and RFN elaborated on the hypothesis, conducted the analyses, helped to interpret the results, and contributed to the writing. LA and MBS funding and elaborated on the hypothesis. MP, GRC, RMY, GH, and KLRB contributed to the panel evaluation and characterization, helped to interpret the results, writing, and discussion. All authors read and approved the final manuscript.

Funding The authors have not disclosed any funding.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Baby B, Antony P, Vijayan R (2018) Antioxidant and anticancer properties of berries. Crit Rev Food Sci Nutr 58:2491–2507. https://doi.org/10.1080/10408398.2017.1329198

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25

Berro I, Lado B, Nalin RS, Quincke M, Gutiérrez L (2019) Training population optimization for genomic selection. Plant Genome 12:190028. https://doi.org/10.3835/plantgenome2019.04.0028

Boyles RE, Ballén-Taborda C, Brown-Guedira G, Costa J, Cowger C, DeWitt N, Griffey CA, Harrison SA et al (2024) Approaching 25 years of progress towards Fusarium head blight resistance in southern soft red winter wheat (*Triticum aestivum* L.). Plant Breed 143:66–81. https://doi.org/10.1111/pbr.13137

Campos GR, Prado M, Borges KLR, Yassue RM, Sabadin F, Silva AV, Barbosa CMA, Sposito MB, Amorim L, Fritsche-Neto R (2023) Construction and genetic characterization of an interspecific raspberry hybrids panel aiming resistance to late leaf rust and adaptation to tropical regions. Sci Rep 13:15216. https://doi.org/10.1038/s41598-023-41728-8

Cohen J (1988) Statistical power analysis for the behavioral sciences. Academic Press, New York

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los CG, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975. https://doi.org/10.1016/j.tplants.2017.08.011

de los Campos G, Sorensen D (2014) On the genomic analysis of data from structured populations. J Anim Breed Genet 131:163–164. https://doi.org/10.1111/jbg.12091

Desta ZA, Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601. https://doi.org/10.1016/j.tplants.2014.05.006

Dias HB, Alvares CA, Sentelhas PC (2017) A century of meteorological data in Piracicaba, SP: Climate changes according to the Köppen classification. In: Brazilian

Congress of Agrometeorology, Symposium on Climate Change and Desertification of the Brazilian Semiarid.

Dias MG, Ribeiro RR, Barbosa A, Jesus CM, Spósito MB (2022) Diagrammatic scale for improved late leaf rust severity assessments in raspberry leaves. Can J Plant Path 45(2):140–147. https://doi.org/10.1080/07060661.2022.2147587

DoVale JC, Carvalho HF, Sabadin F et al (2022) Genotyping marker density and prediction models effects in long-term breeding schemes of cross-pollinated crops. Theor Appl Genet 135:4523–4539. https://doi.org/10.1007/s00122-022-04236-3

Ellis MA, Converse RH, Williams RN, Williamson B (1991) Compendium of raspberry and blackberry diseases and insects, 2nd edn. APS Press, St. Paul

Fernandez GE, Molina-Bravo R, Takeda F (2018) What we know about heat stress in rubus. In: Raspberry: breeding, challenges and advances, pp 29–40

Fritsche-Neto R, Resende MDV, Miranda GV, DoVale JC (2012) Seleção genômica ampla e novos métodos de melhoramento do milho. Revista Ceres 59:794–802. https://doi.org/10.1590/S0034-737X2012000600009

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ et al (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9(2):e90346. https://doi.org/10.1371/journal.pone.0090346

Granato ISC, Galli G, de Oliveira Couto EG, Souza MB, Mendonça LF, Fritsche-Neto R (2018) snpReady: a tool to assist breeders in genomic analysis. Mol Breeding 38:102. https://doi.org/10.1007/s11032-018-0844-8

Hall HK, Hummer KE, Jamieson AR, Jennings SN, Weber CA (2009) Plant breeding reviews. Wiley-Blackwell, New Jersey

Hastie T, Tibshirani R, Narasimhan B, Chu G (2022) Impute: Imputation for microarray data. R package version 1.70.0

Hoque A, Anderson JV, Rahman M (2024) Genomic prediction for agronomic traits in a diverse Flax (*Linum usitatissimum* L.) germplasm collection. Sci Rep 14:3196. https://doi.org/10.1038/s41598-024-53462-w

Isidro J, Jannink JL, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158. https://doi.org/10.1007/s00122-014-2418-4

Iwata H, Minamikawa MF, Kajiya-Kanegae H, Ishimori M, Hayashi T (2016) Genomics-assisted breeding in fruit trees. Breed Sci 66:100–115. https://doi.org/10.1270/jsbbs.66.100

Kainer D, Lanfear R, Foley WJ, Külheim C (2015) Genomic approaches to selection in outcrossing perennials: focus on essential oil crops. Theor Appl Genet 128:2351–2365. https://doi.org/10.1007/s00122-015-2591-0

Kaler AS, Purcell LC, Beissinger T, Gillman JD (2022) Genomic prediction models for traits differing in heritability for soybean, rice, and maize. BMC Plant Biol 22:87. https://doi.org/10.1186/s12870-022-03479-y

Krishnappa G, Savadi S, Tyagi BS, Singh SK, Mamrutha HM, Kumar S, Mishra CN, Khan H, Gangadhara K, Uday G et al (2021) Integrated genomic selection for rapid improvement of crops. Genomics 113:1070–1086. https://doi.org/10.1016/j.ygeno.2021.02.007

Kwong QB, Ong AL, Teh CK, Chew FT, Tammi M, Mayes S, Kulaveerasingam H, Yeoh SH, Harikrishna JA, Appleton DR (2017) Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis guineensis* Jacq.). Sci Rep 7:2872. https://doi.org/10.1038/s41598-017-02602-6

Lara LAC, Santos MF, Jank L, Chiari L, Vilela MM, Amadeu RR, Dos Santos JPR, Pereira GDS, Zeng ZB, Garcia AAF (2019) Genomic selection with allele dosage in panicum maximum jacq. G3 Bethesda 9:2463–2475. https://doi.org/10.1534/g3.118.200986

Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic selection for forest tree improvement: methods. Achiev Perspect Forests 11:1190. https://doi.org/10.3390/f11111190

Legarra A, Garcia-Baccino CA, Wientjes YCJ, Vitezica ZG (2021) The correlation of substitution effects across populations and generations in the presence of nonadditive functional gene action. Genetics 219:iyab138. https://doi.org/10.1093/genetics/iyab138

Lehermeier C, Schön CC, de Los CG (2015) Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics 201:323–337. https://doi.org/10.1534/genetics.115.177394

MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genom 17:144. https://doi.org/10.1186/s12864-016-2443-6

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genomewide dense marker maps. Genetics 157:1819–1829

Montesinos-López OA, Bentley AR, Saint Pierre C, Crespo-Herrera L, Rebollar-Ruellas L, Valladares-Celis PE, Lillemo M, Montesinos-López A, Crossa J (2023) Efficacy of plant breeding using genomic information. Plant Genome 16(2):e20346. https://doi.org/10.1002/tpg2.20346

Montesinos-López OA, Crespo-Herrera L, Xavier A, Godwa M, Beyene Y, Saint Pierre C, de la Rosa-Santamaria R, Salinas-Ruiz J, Gerard G, Vitale P, Dreisigacker S, Lillemo M, Grignola F, Sarinelli M, Pozzo E, Quiroga M, Montesinos-López A, Crossa J (2024) A marker weighting approach for enhancing within-family accuracy in genomic prediction. G3 Genes Genom Genet 14(2):278. https://doi.org/10.1093/g3journal/jkad278

Nadeau S, Beaulieu J, Gezan SA, Perron M, Bousquet J, Lenz PRN (2023) Increasing genomic prediction accuracy for unphenotyped full-sib families by modeling additive and dominance effects with large datasets in white spruce. Front Plant Sci 14:1137834. https://doi.org/10.3389/fpls.2023.1137834

Olatoye MO, Clark LV, Labonte NR, Dong H, Dwiyanti MS, Anzoua KG, Brummer JE, Ghimire BK, Dzyubenko E, Dzyubenko N, Bagmet L, Sabitov A, Chebukin P, Głowacka K, Heo K, Jin X, Nagano H, Peng J, Yu CY, Yoo JH, Zhao H, Long SP, Yamada T, Sacks EJ, Lipka AE (2020) Training population optimization for genomic

selection in miscanthus. G3 Genes Genom Genet 10(7):2465–2476. https://doi.org/10.1534/g3.120.401402

Pérez P, los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198(2):483–495. https://doi.org/10.1534/genetics

Pinczinger D, von Reth M, Hanke MV, Flachowsky H (2021) Self-incompatibility of raspberry cultivars assessed by SSR markers. Sci. Hortic 288:110384

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7(2):e32253. https://doi.org/10.1371/journal.pone.0032253

Prado M, Silva AV, Campos GR, Borges KLR, Yassue RM, Husein G, Akens FF, Sposito MB, Amorim L, Behrouzi P, Bustos-Korts D, Fritsche-Neto R (2024) Complementary approaches to dissect late leaf rust resistance in an interspecific raspberry population. Genes Genom Genet. https://doi.org/10.1093/g3journal/jkae202

Resende R, Resende M, Silva F, Azevedo C, Dapiaggi M, Soares L, Costa E, Martins R, Faria D, Neves L, Oliveira M, Lima B, Alves R, Lima F, Matrangolo W, Silva-Jr O, Grattapaglia D et al (2017) Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity 119:245–255. https://doi.org/10.1038/hdy.2017.37

Rooney TE, Kunze KH, Sorrells ME (2022) Genome-wide marker effect heterogeneity is associated with a large effect dormancy locus in winter malting barley. Plant Genom 15(4):e20247. https://doi.org/10.1002/tpg2.20247

Rossum BJ, Eeuwijk FA, Boer M, Malosetti M, Bustos-Korts D, Millet E, Paulo J (2023) statgenSTA: single trial analysis (STA) of field trials R Package Version 1 11

Runemark A, Vallejo-Marin M, Meier JI (2019) Eukaryote hybrid genomes. PLoS Genet 15(11):e1008404. https://doi.org/10.1371/journal.pgen.1008404

Tan B, Grattapaglia D, Martins GS et al (2017) Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol 17:110. https://doi.org/10.1186/s12870-017-1059-6

Tan B, Grattapaglia D, Wu HX, Ingvarsson PK (2018) Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. Plant Sci 267:84–93. https://doi.org/10.1016/j.plantsci.2017.11.011

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

Wu PY, Ou JH, Liao CT (2023) Sample size determination for training set optimization in genomic prediction. Theor Appl Genet 136:57. https://doi.org/10.1007/s00122-023-04254-9

Würschum T, Maurer HP, Weissmann S, Hahn V, Leiser WL (2017) Accuracy of within- and among-family genomic prediction in triticale. Plant Breed 136:230–236. https://doi.org/10.1111/pbr.12465

Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, Olsen MS, Wang G, Zhang A (2019) Enhancing genetic gain through genomic selection: from livestock to plants. Plant Commun 1(1):100005. https://doi.org/10.1016/j.xplc.2019.100005

Yassue RM, Sabadin F, Galli G, et al. (2021) CV-α: designing validation sets to increase the precision and enable multiple comparison tests in genomic prediction. Euphytica 217:106. https://doi.org/10.1007/s10681-021-02831-x