Theoretical and Applied Genetics (2024) 137:9
https://doi.org/10.1007/s00122-023-04512-w

ORIGINAL ARTICLE

Using visual scores for genomic prediction of complex traits in breeding programs

Camila Ferreira Azevedo[1,2] · Luis Felipe Ventorim Ferrão[2] · Juliana Benevenuto[2] · Marcos Deon Vilela de Resende[1,3,4] · Moyses Nascimento[1] · Ana Carolina Campana Nascimento[1] · Patricio R. Munoz[2]

Received: 9 July 2023 / Accepted: 21 November 2023

Abstract

Key message **An approach for handling visual scores with potential errors and subjectivity in scores was evaluated in simulated and blueberry recurrent selection breeding schemes to assist breeders in their decision-making.**

Abstract Most genomic prediction methods are based on assumptions of normality due to their simplicity and ease of implementation. However, in plant and animal breeding, continuous traits are often visually scored as categorical traits and analyzed as a Gaussian variable, thus violating the normality assumption, which could affect the prediction of breeding values and the estimation of genetic parameters. In this study, we examined the main challenges of visual scores for genomic prediction and genetic parameter estimation using mixed models, Bayesian, and machine learning methods. We evaluated these approaches using simulated and real breeding data sets. Our contribution in this study is a five-fold demonstration: (i) collecting data using an intermediate number of categories (1–3 and 1–5) is the best strategy, even considering errors associated with visual scores; (ii) Linear Mixed Models and Bayesian Linear Regression are robust to the normality violation, but marginal gains can be achieved when using Bayesian Ordinal Regression Models (BORM) and Random Forest Classification; (iii) genetic parameters are better estimated using BORM; (iv) our conclusions using simulated data are also applicable to real data in autotetraploid blueberry; and (v) a comparison of continuous and categorical phenotypes found that investing in the evaluation of 600–1000 categorical data points with low error, when it is not feasible to collect continuous phenotypes, is a strategy for improving predictive abilities. Our findings suggest the best approaches for effectively using visual scores traits to explore genetic information in breeding programs and highlight the importance of investing in the training of evaluator teams and in high-quality phenotyping.

Introduction

Over the last century, plant and animal breeders have used quantitative genetics to estimate genetic parameters and predict phenotypic traits. These traits are typically modeled as a function of the genetic makeup of plants (genotype) and the conditions in which that plant developed (environment). The traditional statistical framework in this field relies on a core assumption, the normality of the residuals and, consequently, the response variables. The use of linear models for phenotypes that follow a Normal (or Gaussian) distribution is attractive due to its simplicity, robustness, straightforward implementation, and support by a well-established theory. However, data collection in plant breeding often involves visual scores to simplify assessments and reduce the costs of phenotyping even, when traits are originally normally distributed. In such cases, statistical challenges may arise do

Communicated by Mikko J. Sillanpää.

✉ Patricio R. Munoz
p.munoz@ufl.edu

[1] Statistics Department, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[2] Horticultural Sciences Department, Blueberry Breeding and Genomics Lab, University of Florida, Gainesville, FL, USA

[3] Department of Forestry Engineering, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[4] Embrapa Café, Brasília, Distrito Federal, Brazil

Published online: 15 December 2023

not exist in the original scale. A central question faced by biometricians is how they should model these phenotypes visually scored and if such normality violation may affect estimates of genetic parameters of key interest.

At least in theory, the violation of the normality assumption in these data sets can invalidate the model and affect future decisions, such as leading to highly imprecise estimates (Schielzeth et al. 2020). To address these issues, breeders and biometricians have used various strategies. The simplest approach is ignoring the scale of the recorded data, under the argument that large sample sizes follow the central limit theorem–which states that treatment means have an approximately normal distribution when sample sizes are large enough (Montesinos-López et al. 2017). Another common alternative is transforming the phenotypes, which can stabilize the residual variation, and hence help fulfill the assumptions required by linear modeling. Although both alternatives are popular in the plant breeding literature, empirical evidence shows that statistical inference's accuracy and power can be reduced when shoehorning the data into classical statistical methods (Stroup 2015).

A more formal approach to analyzing categorical traits relies on using a generalized linear model (GLM) (McCullagh and Nelder 1989). In this model, the mean of response is modeled as a function of explanatory variables, and the response variable is assumed to be conditionally distributed according to an exponential family distribution (e.g., Binomial or Poisson distributions). Several methodologies have been proposed in the genetics field to handle categorical data, including mixed-model-based approaches (Harville and Mee 1984; Gilmour et al. 1985), Bayesian methodologies (Montesinos-López et al. 2015a, 2020a; Pérez-Rodríguez et al. 2020), and machine learning and artificial intelligence techniques (González-Recio et al. 2014; Montesinos-López et al. 2020b; Montesinos López et al. 2022a). However, breeders are still seeking answers regarding easily implementable methods that provide accurate results, the best approach to record categorical data instead of recording the trait in its original scale and whether they should continue with this practice or make an effort to record data in the original scale.

Another important aspect for practical implementation is that contemporaneous breeding programs have modeled phenotypic observations as a function of variations at the DNA level. Referred to as genomic selection, this methodology is a form of marker-assisted selection in which all available molecular markers are used to predict quantitative traits (Meuwissen et al. 2001). Despite its importance, the debate around using of non-continuous traits remains the same: most genomic prediction models are based on linear regression models that assume continuous and normally distributed phenotypes, without clear evidence on the impact of normality violation on estimating genetics parameters. Some recent studies have relaxed these assumptions and applied threshold models and Bayesian ordinal regression. For example, Montesinos-López et al. (2015b) introduced genomic selection models for ordinal traits in maize, and reported gains when genotype-by-interaction was taken into account. They also reported the use of ordinal logistic regression for predictions, under the argument that ordinal models are more robust for dealing with outlying data and provide interpretable results (Montesinos-López et al. 2015b). In animal science, ordinal and continuous data were compared, and the use of threshold traits resulted in markedly lower accuracy than a linear model (Kizilkaya et al. 2014). More recently, Merrick et al. (2022) reported that using machine learning methods led to higher predictive accuracy for the classification and prediction of traits with skewed distributions. However, most of these studies have focused on the predictive performance rather than estimating genetic parameters of key interest, such as selection gain and marker effects. They have also not fully addressed the challenges breeders face when assigning scores instead of measuring continuous traits.

In this context, it is still unclear how to analyze phenotypic data that are normally distributed but are categorized. A prime example is yield evaluation in fruit crops. Using Blueberry (*Vaccinium spp.*) as our biological model, yield is commonly evaluated after all berries are manually picked and weighed. Remarkably, harvesting takes place multiple times during the crop season, which makes the process slow, laborious, and expensive. As an alternative, breeders visually classify the genotypes using scores that can range between 1(low production) and 5 (high production). Despite the simplicity, the use of visual assessments does not relieve breeders of important decisions, i.e., the choice of increase costs of screening a large population using visual scores or conducing reduced experimental by using numerical scoring for continuous variation in a variable.

Aiming to shed light on the relevance of using visual score traits in plant breeding, we conceptualized this study in two sections. First, we simulated continuous data with Gaussian distribution, categorized it, and included different levels of noise–mimicking potential errors associated with recording visual scores. We draw attention on the impact of using categorical traits for prediction and genetic inference using parametric and non-parametric models. In the sequence, we used real data collected in a large blueberry breeding population and modeled categorical traits evaluated over multiple years and locations. Collectively, in this study we addressed the following questions: (i) what is the best strategy for recording and modeling categorical data? (ii) what is the effect of the operators' experience (error level) when estimating genetic parameters? (iii) should categorical traits be modeled using parametric or non-parametric methods? and finally, (iv) how can breeders allocate resources for

phenotyping to collect continuous and categorical data, to maximize predictive gains?

Material and methods

The Materials and Methods section of the study is organized as follows. The "Statistical and machine learning methods" section outlines the use of parametric and non-parametric models for predicting and estimating genetic parameters. The "Simulated Data" and "Scenarios of Analyzes" sections describe the development of simulations for the experiments, using a stochastic model that considers various levels of noise/errors and the scenarios of Analyzes evaluated using the simulated data. The "Real Data" section applies the findings from the simulated results to real data Analyzes on blueberries. Finally, the "Genomic prediction and efficient measures calculation" section presents the metrics used to compare the different analysis approaches.

Statistical and machine learning methods

For data modeling, we employed two main approaches: statistical methods (including Generalized Linear Mixed Model, Linear Mixed Model, Bayesian Ordinal Regression Model, and Bayesian Linear Regression Model) and machine learning methods (Random Forest Regression and Random Forest Classification).

Generalized linear mixed model and linear mixed model

The Generalized Linear Mixed Model (GLMM) is defined as:

$$\boldsymbol{l} = \mathbf{1}\mu + \boldsymbol{Zu} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{l}$ is the vector of latent variables (called liabilities) for the vector of phenotypic values represented by $\boldsymbol{y}$, $\mathbf{1}$ is a vector of the same dimension of $\boldsymbol{l}$ being all elements equal to 1, μ is the trait mean, $\boldsymbol{u}$ is the vector of additive genetic random effects of individuals with $\boldsymbol{u}|\sigma_u^2 \sim N(0, \sigma_u^2\boldsymbol{G})$, $\mathbf{G}$ is the additive genomic relationship matrix (VanRaden 2008) between individuals and σ_u^2 is the additive genetic variance, $\boldsymbol{Z}$ is the matrix of incidence of random effects and $\boldsymbol{e}$ is the vector of random errors. In a generalized linear model, it is assumed that each observation of the variable Y has a distribution belonging to the exponential family. In this case, the expected value of Y is defined as:

$$E(Y) = g^{-1}(l) \tag{2}$$

where $g(.)$ is the link function. The linear mixed model (LMM) is a particular case of GLMM, in which the variable Y follows a normal distribution with mean $\mathbf{1}\mu + \boldsymbol{Zu}$ and a covariance matrix given by $\boldsymbol{I}\sigma_e^2$, where σ_e^2 is the residual variance, and whose link function is the identity (McCullagh and Nelder 1989). Suppose the Y variable consists of levels of some categorical factors and has a natural ordering, an adequate link function is the *probit* function. Ordinal categorical predictions for phenotypes with K categories are defined based on threshold parameters $\boldsymbol{\gamma}' = (\gamma_0 = -\infty, \gamma_1, \gamma_2, \ldots, \gamma_{K-1}, \gamma_K = +\infty)$ that have a continuous scale and relate to the observed ordinal categorical response, according to:

$$Y_{Ki} = \begin{cases} 1 & if \quad \gamma_0 < l_i \leq \gamma_1 \\ 2 & if \quad \gamma_1 < l_i \leq \gamma_2 \\ \vdots & \\ K & if \quad \gamma_{K-1} < l_i \leq \gamma_K \end{cases}$$

where $i = 1, \ldots, n$ and n is the number of the observations.

Therefore, in the ordinal model, Y_i is a random variable that takes valuesl, ..., K, with the following probabilities:

$$\begin{aligned} &P\left(Y_i = k|\mu, \boldsymbol{u}, \boldsymbol{\gamma}\right) = P\left(\gamma_{k-1} < l_i \leq \gamma_k|\mu, \boldsymbol{u}, \boldsymbol{\gamma}\right) \\ &= F_N(\gamma_k - \mu - u_i) - F_N(\gamma_{k-1} - \mu - u_i) \end{aligned} \tag{3}$$

where F_N is the Gaussian cumulative distribution function.

Nelder and Wedderburn (1972) demonstrated the transformation of $\widetilde{\boldsymbol{y}}$ given by the equation:

$$\widetilde{\boldsymbol{y}} = \mathbf{1}\mu + \boldsymbol{Zu} + g'(E(Y))(\boldsymbol{y} - E(Y)) \tag{4}$$

where $\widetilde{\boldsymbol{y}}$ is a linear combination of the expected value of the liability and the discrepancy between the observed and adjusted values, and $g^{'}(.)$ is the first derivative of $g(.)$. Following this transformation, $g'(E(Y))(\boldsymbol{y} - E(Y))$ follows a normal distribution. In this case, Henderson's generalized linear mixed model equations (GLMME) can lead to the BLUP in linear mixed models when the response variable is defined as the latent variable $\widetilde{\boldsymbol{y}}$.

So, the estimation and prediction algorithms for the linear mixed model case can be adapted as the mixed model equations below (Resende et al. 2018):

$$\begin{bmatrix} \mathbf{1}'\boldsymbol{S}^{-1}\mathbf{1} & \mathbf{1}'\boldsymbol{S}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{S}^{-1}\mathbf{1} & \boldsymbol{Z}'\boldsymbol{S}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\frac{1}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \widehat{\mu} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\boldsymbol{S}^{-1}\widetilde{\boldsymbol{y}} \\ \boldsymbol{Z}'\boldsymbol{S}^{-1}\widetilde{\boldsymbol{y}} \end{bmatrix} \tag{5}$$

where $\boldsymbol{S}^{-1}$ is a diagonal matrix with elements equal to $V(Y)\sigma_e^2$ and σ_e^2 is the residual variance on the continuous scale (liabilities). The Restricted maximum likelihood (REML) estimators are given by:

$$\widehat{\sigma}_u^2 = \frac{\widehat{\mu}^2\boldsymbol{G}^{-1}}{q - tr(\boldsymbol{G}^{-1}C^{22})/\sigma_u^2}$$

$$\widehat{\sigma}_e^2 = \frac{(\boldsymbol{y} - \mathbf{1}\widehat{\mu} - \boldsymbol{Z}\widehat{\boldsymbol{u}})'\boldsymbol{S}^{-1}(\boldsymbol{y} - \mathbf{1}\widehat{\mu} - \boldsymbol{Z}\widehat{\boldsymbol{u}})}{n - 1 - q - tr(\boldsymbol{G}^{-1}C^{22})/\sigma_u^2}\sigma_e^2$$

where n is the number of the observations, q is the number of random effect levels, tr is trace matrix operator, σ_e^2 and σ_u^2 are values obtained in the previous iteration of the algorithm and C^{22} is the partition of the inverse of the coefficients of the mixed model equations, referring to random effects.

And the solution to the GLMME may be expressed in the following way:

$$\widehat{\mu} = \left(\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{V}^{-1}\widetilde{\boldsymbol{y}} \tag{6}$$

$$\widehat{\boldsymbol{u}} = \sigma_u^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\widetilde{\boldsymbol{y}} - \mathbf{1}\mu) \tag{7}$$

where $\boldsymbol{V} = \boldsymbol{S} + \sigma_u^2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$.

These procedures were implemented in the ASReml package in software *R* (Butler 2022), for each replicated data set within the scenarios and real data set.

Bayesian ordinal regression model and Bayesian linear regression model

Model (1) can be used under a Bayesian framework. The Bayesian Ordinal Regression Model (BORM) model assumes the following prior distribution for the unknown parameter vector $\boldsymbol{\theta} = (\mu, \boldsymbol{u}, \boldsymbol{l}, \boldsymbol{\gamma}, \sigma_u^2)$:

$$\mu \sim N(0, \sigma_\mu^2)$$

$$\boldsymbol{u}|\sigma_u^2 \sim N(\mathbf{0}, \sigma_u^2\boldsymbol{G})$$

$$\sigma_u^2 \sim \chi^{-2}(v, s^2)$$

$$\gamma_k \sim U(a_k, b_k)$$

$$\boldsymbol{l}|\mu, \boldsymbol{u} \sim N(\mathbf{1}\mu + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{I}\sigma_e^2) \tag{8}$$

where $\boldsymbol{I}$ is an identity matrix and $\sigma_e^2 = 1$ to reach identifiability for unobservable liabilities (even when the number of unknown parameters is higher than the sample size), σ_μ^2 is a value assumed as 10^{10} that represents a vague prior knowledge, $\mathbf{G}$ is the additive genomic relationship matrix (VanRaden 2008) between individuals and σ_u^2 is the additive genetic variance. The v, s^2, a_k and b_k are called hyperparameters.

The joint posterior density of $\mu, \boldsymbol{u}, \boldsymbol{\gamma}, \sigma_u^2$ and the liabilities $\boldsymbol{l}$ is given by:

$$p\left(\mu, \boldsymbol{u}, \boldsymbol{l}, \boldsymbol{\gamma}, \sigma_u^2|\boldsymbol{y}\right) \propto p(\boldsymbol{y}|\boldsymbol{l}, \boldsymbol{\gamma})p(\boldsymbol{l}|\mu, \boldsymbol{u})p(\boldsymbol{\gamma})p(\mu)p\left(\boldsymbol{u}|\sigma_u^2\right)p\left(\sigma_u^2\right) \tag{9}$$

The inference of the parameters $(\mu, \boldsymbol{u}, \boldsymbol{l}, \boldsymbol{\gamma}, \sigma_u^2)$ is based on their marginal posterior distributions, obtained indirectly from full conditional posterior distributions through the Markov Chain Monte Carlo (MCMC) algorithms. These full conditional posterior distributions were presented by Montesinos López et al. (2022b). These procedures were implemented by the BGLR package in software *R* (Pérez and de los Campos 2014) to each replicated data set within each scenario and to the real data by defining 500,000 iterations for the MCMC algorithms, a burn-in period of 50,000 MCMC cycles, and thin equals to 10 before saving samples from each, totaling 45,000 MCMC cycles. The $\widehat{\boldsymbol{u}}$, $\widehat{\mu}$, $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_e^2$ estimates were obtained as the posterior mean of their respective marginal posterior distributions. In the Bayesian Linear Regression Model (BLRM), the joint posterior density is simplified to $p\left(\mu, \boldsymbol{u}, \sigma_u^2, \sigma_e^2|\boldsymbol{y}\right) \propto p(\boldsymbol{y}|\mu, \boldsymbol{u})p(\mu)p\left(\boldsymbol{u}|\sigma_u^2\right)p\left(\sigma_u^2\right)p\left(\sigma_e^2\right)$ because the own response variable Y assumes the Normal distribution, e.g.,, $\boldsymbol{y}|\mu, \boldsymbol{u}, \sigma_e^2 \sim N(\mathbf{1}\mu + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{I}\sigma_e^2)$.

Random forest regression and random forest classification

Random Forest Regression (RFR) and Random Forest Classification (RFC) are supervised machine learning methods based on tree algorithms that can apply to continuous, binary and categorical variables, respectively (James et al. 2013). The tree algorithms divide the predictor space ($X_1, X_2, \ldots, X_p$ –in this study, molecular markers) into several non-overlapping regions ($R_1, ..., R_J$), and these stratifications are based on the optimization of cost functions. The regression tree is indicated for continuous traits, and the goal is to find boxes $R_1, ..., R_J$ that minimize the Residual Sum of Squares given by:

$$\sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i - \widehat{y}_{R_j}\right)^2$$

where $\widehat{y}_{R_j}$ is the mean of response observations (continuous phenotypes) within the jth region. However, instead of considering each possible partition of space in J regions to reduce the computational time of the Analyzes, recursive binary splitting is performed by selecting the predictor X_j and the cutpoint s and then minimizing the equation:

$$\sum_{i:x_i\in R_1(j,s)}\left(y_i - \widehat{y}_{R_1}\right)^2 + \sum_{i:x_i\in R_2(j,s)}\left(y_i - \widehat{y}_{R_2}\right)^2 \tag{10}$$

where $\widehat{y}_{R_1}$ and $\widehat{y}_{R_2}$ are, respectively, the mean of response observations in $R_1(j, s)$ and $R_2(j, s)$.

The classification tree procedure is very similar to a regression tree, but is indicated for binary and categorical

traits, and uses other cost functions like the Gini index, given by:

$$G = \sum_{k=1}^{K} \widehat{p}_{jk}(1 - \widehat{p}_{jk})$$

where $\widehat{p}_{jk}$ is the proportion of observations in the jth region that are from the kth category. Typically, a single regression or classification tree has high variability in its predictions and, as a result, a reduced ability to make accurate predictions. To improve the predictive and classification performance of the tree, refinements such as Random Forest can be used. Random Forest builds multiple trees that are decorrelated by using a subset of predictor variables in each partition, and then averages or takes the mode of the predicted values. This results in independent predicted values, which reduces the variability of the tree (Ho 1995). The number of predictors (m) suggested by Hastie et al. (2009) is $m = \sqrt{p}$ for classification trees (RFC) and $m = p/3$ for regression trees (RFR).

These procedures were implemented by the randomForest package in software *R* (Liaw and Wiener 2022) to each replicated data set within each scenario and real data.

Simulated data

The effect of categorizing continuous traits using a simulated data set was first studied by simulating phenotypic traits with contrasting genetic architectures. The simulated genome consisted of 10 pairs of chromosomes with a genetic length of 1.43 Morgans and a physical length of 8×10^8 base pairs. The recombination rate was calculated based on the genome size (i.e., 1.43 Morgans per 8×10^8 base pairs, which equals 1.8×10^{-9} per base pair), and the mutation rate was set to 2×10^{-9} per base pair (Batista et al. 2021). Sequences for each chromosome were randomly chosen to have 1,000 causal loci per chromosome (a total of 10,000 across the genome) and were generated using the Markovian Coalescent Simulator (Chen et al. 2009). After generating genome sequences, we created 50 founder genotypes that were used as initial parents in the burn-in phase. The following steps involved crosses between highly heterozygous hybrids. After simulating the crosses, 10% of the resulting F1 progenies (5,000 genotypes) were selected based on their estimated breeding value (500 genotypes). Then, using genomic models, we selected 50 genotypes to be parents in the next breeding cycle. This simulation represents a typical small effective population size (Ne = 50). The methods and analysis scenarios were evaluated using the average of ten replicates, and each replicate consisted of: (i) a burn-in phase that consisted of 20 cycles of breeding, and (ii) an evaluation phase that simulated future breeding with different analysis strategies. In this way, we simulated a classical recurrent selection breeding program in which the allele frequencies of target traits are increased by selecting the best individuals and crossing them to create a new generation.

In our breeding design, we simulated traits with two genetic architectures within the current population: (i) a qualitative trait controlled by three large quantitative trait loci (QTL) and high heritability (0.60); and (ii) a quantitative trait following the infinitesimal model, with 100 QTL and low heritability (0.10). For each QTL, we assigned one additive effect on the phenotype following a normal distribution with zero mean and variance, resulting in the desired heritability level (Gaynor et al. 2021). We added a random deviation from a normal distribution $N(0,100)$ to the genotypic value. All simulations were carried out using AlphaSimR (Gaynor et al. 2021), following a similar crossing and selection design as those described by Batista et al. (2021).

As part of the simulation process, continuous phenotypic traits (Y_i) were categorized as follows: two (1–2), three (1–3), five (1–5), and nine (1–9) categories. The use of four different numbers of categories mimics visual scores which are commonly used in plant breeding programs. To create these categories (Y_{K_i}), we used the following thresholds to categorize the simulated phenotype values:

$$Y_{K_i} = \begin{cases} 1 & if \quad -\infty < Y_i \leq \tau_1 \\ 2 & if \quad \tau_1 < Y_i \leq \tau_2 \\ \vdots & \\ K & if \quad \tau_{K-1} < Y_i \leq \tau_K \end{cases}$$

where $\tau_1, \tau_2, \ldots, \tau_K$ are the non-equidistant thresholds based on quantiles distribution and K is the number of categories ($K = 2, 3, 5 \, or \, 9$).

Assuming that visual scores may be subject to errors due to the experience of the person recording them, we also simulated different levels of noise, or error. We considered three levels of error: low, moderate, and high, which corresponded to the introduction of 20%, 50%, and 70% error variance into the scores, respectively. This approach is similar to what has been reported in genomic studies that have evaluated the effect of mislabeling on genomic prediction trials (Biffani et al. 2017; Yabe et al. 2018). Importantly, an error rate of 20% –even considered as "low" in this study–is still a very conservative value for most breeding programs. Figure 1 summarizes the different categorical classes and noise levels.

Scenarios of analysis

For statistical purposes, genomic prediction was performed according to three theoretical scenarios:

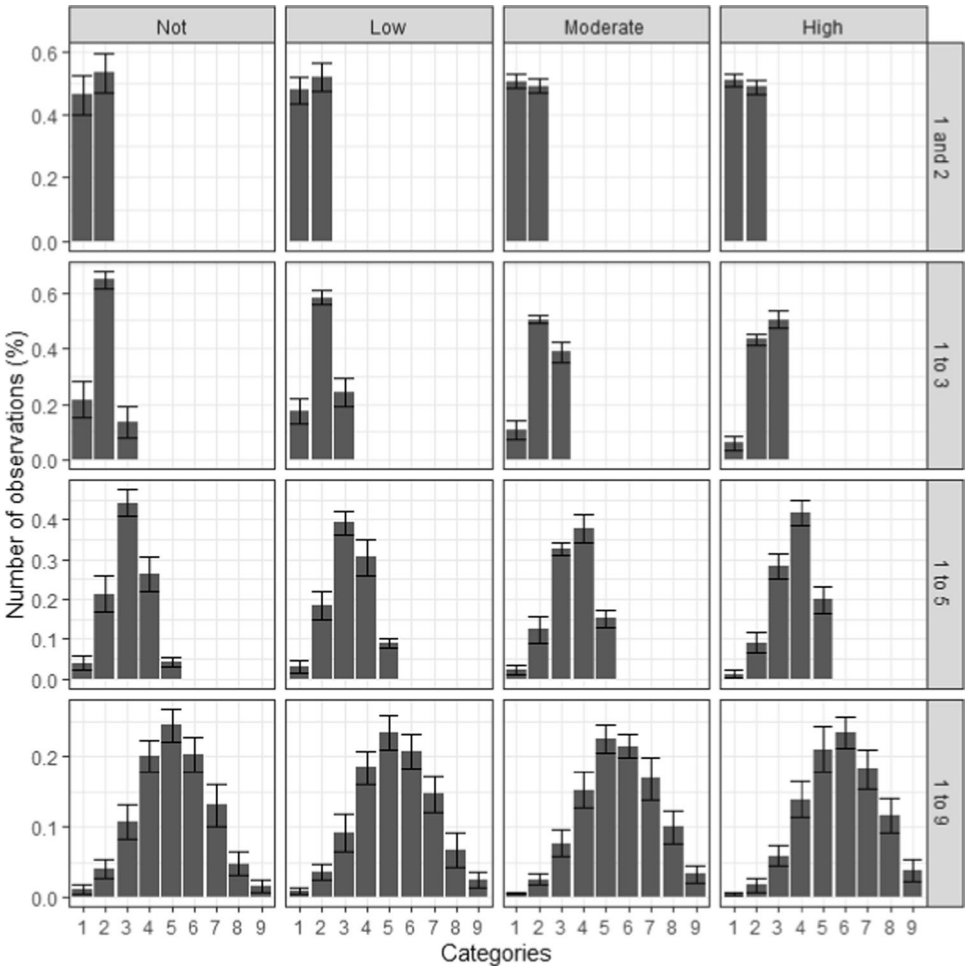

Fig. 1 Distribution of observations simulated by four different categories, considering three noise levels (low – 20% of misclassification, moderate – 50% of misclassification, and high – 70% of misclassification) and no errors. Error bars serve as graphical representations of the variability in simulated data across these scenarios

- Continuous (CONT—benchmark): The simulated phenotypes (Y_i) were used as the response variable in the Linear Mixed Model, Bayesian Linear Regression Model, or Random Forest Regression. This represents the benchmark scenario, where the true continuous distribution of the phenotypes is considered and analyzed with an appropriate model and probability distribution.
- Categorical-Continuous (CAT-CONT—practical): After categorizing the continuous phenotypes (Y_{K_i}), we decided to process the data into classical statistical methods and used the ordinal response variable in the Linear Mixed Model, Bayesian Linear Regression Model or Random Forest Regression. This scenario represents what is typically done in breeding programs, where the phenotypes are collected as categorical but analyzed as continuous.
- Categorical (CAT—formal): After categorizing the continuous phenotypes (Y_{K_i}), we used the categorical response variable into a Generalized Linear Mixed Model, Bayesian Ordinal Regression Model or Random Forest Classification. This scenario represents a formal statistical procedure for Analyzes of categorical phenotypes.

For a graphic representation of the "Simulated Data", "Statistical and machine learning methods" and "Scenarios of Analyzes" sections, see Fig. 2.

Real data

We extended our simulated Analyzes to real data using blueberry as our biological model. Briefly, as part of the recurrent selection strategy, the UF blueberry breeding program annually makes up to 150–200 crosses. These crosses are designed based on a combination of phenotypic, pedigree, and molecular information that predicts plant performance for yield, fruit quality, flavor, disease/pest resistance, and early season production. From crosses to a cultivar, a four-stage selection approach is used. In the first stage, 20,000 progenies are planted in high-density plantations from the approximately 200 crosses. The first evaluation cycle is conducted on one-year-old seedlings (Stage I), and about 10% of seedlings are advanced to the second stage (Stage II). In the second year, with more fruits available for evaluation, a new selection (10% of the approximately 2,000 remaining plants) is performed. Stage III consists of clonally propagated plants

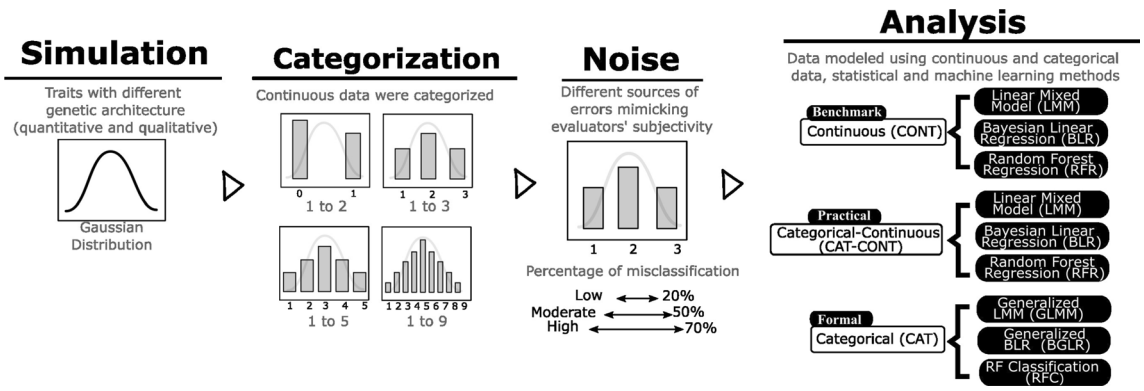

Fig. 2 The following flowchart illustrates the procedures used in this study. The simulation component involves simulating traits with two genetic architectures (quantitative and qualitative) and Gaussian distributions. The categorization component involves delineating the continuous phenotypic traits into categories (1 to 2, 1–3, 1–5, and 1–9). The noise component includes the creation of noise levels (low–20% of misclassification, moderate–50% of misclassification, and high–70% of misclassification) in the categorization, for mimicking the evaluator experience. The analysis component represents the application of parametric (Generalized and Linear Mixed Models, Bayesian Ordinal and Bayesian Linear Regression Models) and non-parametric (Random Forests Regression and Classification) methods to continuous and categorical traits. The CONT analysis scenario represents the benchmark scenario when continuous phenotypes are the response variables in the LMM, BLRM, and RFR. The CAT-CONT analysis scenario represents the typical practice in breeding programs where categorical phenotypes are the response variables in the LMM, BLRM, and RFR. The CAT analysis scenario represents a formal statistical procedure where the categorical phenotypes are the response variable in the GLMM, BORM, and RFC

that have been established in a commercial field and evaluated in a 15-plant clonal plot. Of the genotypes in Stage III, approximately 10% of the most promising are selected to move on to Stage IV. At this stage, evaluations are conducted on commercial farms throughout the state of Florida, using clonal field plots with 5–45 plants per plot. The final selection consists of genotypes that perform well across years and locations for evergreen or deciduous systems and are released as cultivars. During this four-stage selection process, genotypes have been visually evaluated using 1 (low) to 5 (high) categorical scores for yield, vigor, healthiness, yield estimated via floral buds, yield estimated via flower numbers, and yield estimated via green berries. Firmness and size have also been visually scored on a scale of 1 (low) to 3 (high). Table 1 shows the average number of samples genotyped and phenotyped across various locations and seasons.

For the genomic analyzes, we followed the same approach as described in Ferrão et al. (2021) and Benevenuto et al. (2019). Briefly, genotyping was performed using the "Capture-Seq" approach, and reads were aligned against the largest scaffolds of each of the 12 homoeologous groups of *Vaccinium corymbosum* cv. "Draper" genome assembly (Colle et al. 2019). SNPs were called with FreeBayes v.1.3.2, using 10,000 probe positions as targets (Garrison and Marth 2012). Loci were filtered, applying the following criteria: (i) minimum mapping quality of 10; (ii) only biallelic locus; (iii) maximum missing data of 50%; (iv) minor allele frequency of 1%; and (v) minimum and maximum mean sequence depth of 3 and 750 across individuals,

Table 1 Mean of the number of observations of eight categorical traits, evaluated over several seasons and breeding stages of the blueberry breeding program, in four Florida regions

Trait	Category	North	Citra[1]	Central	South	Seasons	Breeding stages
Yield	1–5	713	931	67	85	2014,2015,2020,2021	II, III, IV
Vigor	1–5	596	119	89	71	2020,2021,2022	III, IV
Health	1–5	590	114	85	68	2020,2021,2022	III, IV
Yield estimate floral Buds	1–5	323	123	89	75	2021,2022	III, IV
Yield estimate flowers	1–5	183	70	89	70	2021,2022	III, IV
Yield estimate green Berries	1–5	156	–	78	60	2021,2022	III, IV
Firmness	1–3	–	–	76	58	2022	III, IV
Size	1–3	–	–	76	56	2022	III, IV

[1]The "Plant Science Research and Education Unit" at the University of Florida is located in Citra, Florida

respectively. A total of 63,552 SNPs were kept after these filtering steps. Sequencing read counts per allele per individual were extracted from the variant call file using vcftools v.0.1.16 (Danecek et al. 2011). And were used as input to estimate the allele dosage, according to the "norm model" in the updog 2.1.0 R package.

Genomic prediction and efficient measures calculation

In the real data analysis, a single genomic prediction model was adjusted for each trait, location and season using mixed models, Bayesian and machine learning methods. These models were identical to those used in the simulations, with the exception of incorporating the age of the plant as a fixed effect in the statistical methods and as a factor in the random forest model.

We compared the prediction performances of methods in the simulated data using fivefold cross-validation and the following metrics: (i) accuracy, which is the correlation between the true breeding value (TBV) and genomic estimated breeding values (GEBV); (ii) prediction bias, which is the deviation from one of the coefficient of regression between TBV and GEBV; (iii) the percentage of agreement between the top 10% of individuals selected by each approach (CONT, CAT-CONT and CAT) and true breeding value; and finally (iv) the correlation between the marker effects estimated by each approach. In the real data, we used the same cross-validation approach and the predictive ability to compare the methods, which is the correlation between the phenotype and GEBV.

We also computed genetic parameters in terms of heritability to both data sets and selection gain to simulated data. For the Bayesian approaches, heritability was computed for the kth value of the Markov chain of heritability, and is given by: $h^{2(k)} = \frac{\sigma_u^{2(k)}}{\sigma_u^{2(k)}+\sigma_e^{2(k)}}$, where $\sigma_u^{2(k)}$ and $\sigma_e^{2(k)}$ are values of the variance components of the kth iteration of the MCMC algorithm. Subsequently, the posterior mean was calculated. For the Henderson's equations approach, the heritability is given by:$h^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$, where σ_u^2 is estimated using REML, $\sigma_e^2 = 1$ in the GLMM, and σ_e^2 is estimated using REML in the LMM. For the random forest approach, the variances were calculated as being the variance of the GEBV and residuals vector and then the heritability values were obtained by $h^2 = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$.

Selection gain was calculated using the following expression, considering the selection of 10% of individuals by each approach: $GS = \left(\overline{X}_s - \overline{X}_o\right)h^2$, where $\overline{X}_s$ is the mean TBV of the selected population, $\overline{X}_o$ is the mean TBV of the original population, and h^2 is the heritability estimate.

Results

Prediction accuracy of different combinations of categorical levels and errors

In breeding programs, phenotypic traits are often recorded using visual scores traits. Herein, we first investigated the impact of categorizing traits that are continuous by nature and tested different combinations of categorical levels and error (Fig. 3). When comparing both extremes, the use of binary categories was a poor simplification of continuous scenarios (Fig. 3a); and the use of more categories (1–9) was noisier (Fig. 3d). In general, we found that an intermediate number of categories (such as 1–3 and 1–5) showed a good compromise in terms of predictive accuracy and error, above the other categories classification tested (Fig. 3b, c).

Statistical models for categorical traits

After discussing the relevance of using different numbers of categories to simplify continuous traits, herein we addressed how breeders should statistically model these traits that are continuous by nature, but they are visually categorized for simplicity. In the Fig. 3, we show three scenarios of accuracy evaluation, where breeders need to deal with a combination of strategies, that is, analyzing visual categorical data with LMM and BLRM (CAT-CONT) and modeling categorical traits either via GLMM or BORM (CAT). The connection between modeling strategy and genetic architecture is a second piece of relevant information. We observed that traits simulated with quantitative and qualitative nature produced very similar results (Fig. 3, Supplemental Fig. 1). By focusing primarily on the quantitative results, BLRM and LMM showed the highest accuracies for continuous scale phenotypes (0.69 and 0.73, respectively). When using standard "normal" models to predict a categorical trait (CAT-CONT), we observed reasonable accuracy values (Fig. 3).

While LMM and BLRM may produce consistent results, they may not be the best approach in terms of accuracy. First, we note that whenever simplification is performed by transforming continuous data into categorical data, there are losses in accuracy. A second important aspect relies on the relevance of including noise in the data Analyzes. We observed that RFC was less sensitive to noise errors, resulting in satisfactory predictive accuracy. On the relevance of using non-parametric methods, RFC had a better predictive performance than RFR, even in scenarios with a larger number of categories. Therefore, we emphasize that when categorical traits are presented, they should be framed as a classification problem, even if the dataset presents an approximately normal distribution.

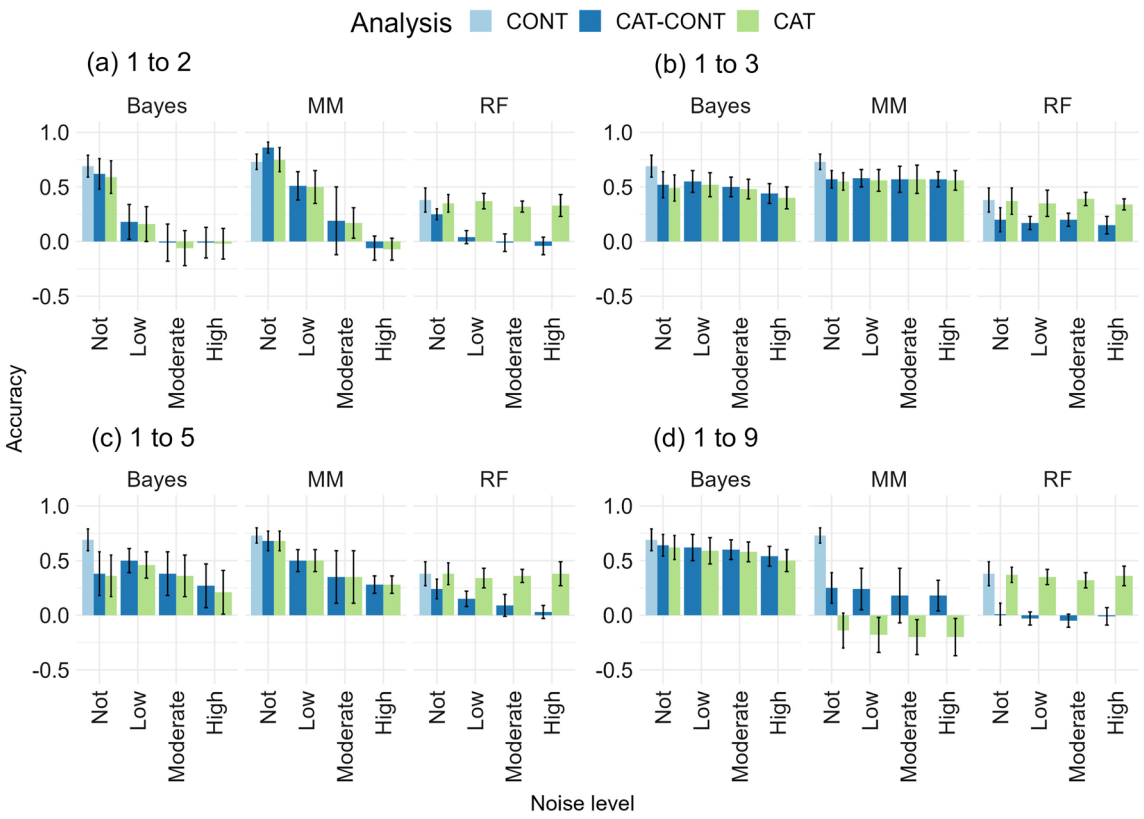

Fig. 3 Accuracy of different methods (Bayesian Ordinal and Bayesian Linear Regression Models (Bayes), Generalized and Linear Mixed Models (MM), and Random Forests Regression and Classification (RF)) under cross-validation procedures for simulated categorical traits, with different category levels (a—1 to 2; b—1 to 3; c—1 to 5; d—1 to 9), under different levels of noise (low—20% misclassification, moderate—50% misclassification, and high—70% misclassification) and no errors, relative to continuous traits. The traits had a quantitative genetic architecture, with 100 QTLs and heritability equal to 0.10

Estimating genetic parameters

Also relevant, breeders rely on the inference of genetic parameters to assist future decisions. When we computed the Person's correlation between marker effects estimated in each scenario (below the diagonal, Fig. 4; Supplemental Figs. 2—5), for categories ranging from 1 to 5 the correlation values were moderate to high. We also observed a large percentage of agreement between the top 10% of individuals selected across the scenarios (above the diagonal, Fig. 4; Supplemental Figs. 2—5). Importantly, Bayesian models reported the highest percentage of agreement on selecting the top 10% of individuals, compared to the true breeding value (first row in the matrices, Fig. 4; and Supplemental Figs. 2—5). Another central genetic parameter is heritability. We found that on the continuous distribution, Bayesian and mixed models recovered the simulated value of heritability (Fig. 5, Supplemental Figs. 6), suggesting that our simulation was appropriate. When contrasting different scenarios, all methods underestimated heritability values. As an important trend, the Bayesian ordinal regression (CAT) presented more robust results in recovering the estimated value. On the other hand, the use of RF did not result in biased estimates (Supplemental Figs. 7, 8). Finally, we investigated the impact on genetic gain, which relies primarily on heritability values (Supplemental Figs. 9, 10). We found a severe underestimation of genetic gains.

Genomic prediction using real blueberry data

The prediction results detected in the simulated data encouraged us to explore genomic prediction in blueberries from three different angles. Importantly, the present study encompasses the largest breeding population used up to date to dissect the importance of using genomic information to predict traits visually scored in the fruit literature. To this end, we first computed predictive ability and genetic parameters. The use of Bayesian model showed the best results (Fig. 6 and Supplemental Table 1). In the sequence, we explored a practical decision underlying population training size. Namely, by integrating categorical and continuous data, we identified a group of 179 samples that were both phenotyped for yield in a continuous (kg per bush) and visually scored (1–5 scores). The estimated heritability using continuous data

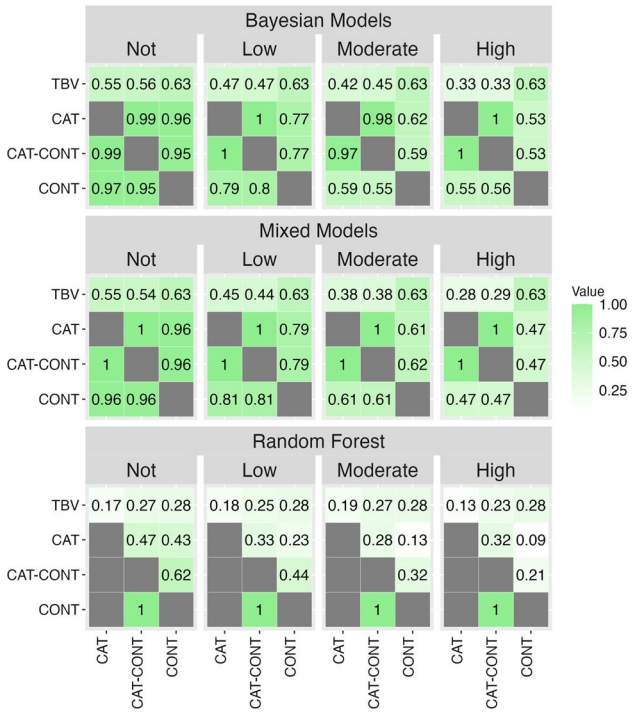

Fig. 4 Percentage of agreement between 10% of individuals selected (above the diagonal) for 1–5 simulated categorical traits, and correlation between marker effects estimated (below the diagonal) by various analysis approaches (CONT, CAT-CONT and CAT) using Bayesian Ordinal and Bayesian Linear Regression Models, Generalized and Linear Mixed Models, and Random Forests Regression and Classification methods, under different levels of noise (low–20% of misclassification, moderate–50% of misclassification, and high–70% of misclassification) and no errors. The traits had a quantitative genetic architecture, with 100 QTLs and heritability equal to 0.10. The first row in the matrices represents the percentage of agreement between the 10% of individuals selected by the approaches and the true breeding values (TBV)

was 0.28, and a similar value was computed for the visual scores (Supplemental Table 1). However, substantial differences were observed for prediction. Yield prediction for the 179 genotypes using continuous data and cross-validation scheme was 0.17. To test the relevance of leveraging predictive ability at the cost of including more visually scored individuals, we trained our model using a new set of 2,323 genotypes–all visually scored using categories ranging from 1 to 5. This new set of genotypes was used to predict the 179 genotypes collected at the continuous distribution. Remarkably, predictive ability increased to 0.35, representing more than a 100% gain over the continuous metric. This fact leads to us raising our last scientific question: what is the best alternative to combine continuous and categorical data in a single framework for genomic prediction?

To answer this question, we used the same stochastic process to simulate traits but under four different genetic architectures (Fig. 7). By mimicking our breeding program, we simulated 200 phenotypes in a continuous distribution. Additionally, new genotypes (ranging from 0 to 4,800 with an increment in 200 individuals) were categorized with different error levels. Herein, our benchmark is the predictive ability computed using the 5,000 continuous phenotypes. Systematically, we included phenotypes collected at a categorical distribution, and checked the predictive ability (Fig. 7). As might be expected, genetic architecture and noise levels are the main drivers of predictive ability, with simple genetic architecture and low levels of errors leading to higher predictive ability. When no errors are simulated, increasing the population size by using more categorical traits will always increase the predictive ability of continuous traits. This is an ideal but unrealistic scenario; a more realistic approach is a program operating with certain noise levels. A more realistic approach is a breeding program operating with a low noise level. Herein, including more categorial phenotypes only add to predictive ability when error levels are low, a fact that sheds light on the importance of properly training human resources for data collection.

Discussion

In this study, we discussed potential scenarios of analyzing Gaussian traits, that are visually scored. Considered as a common practice in multiple breeding programs, we drew attention on the relevance of using different statistical models, forms of data collection, and impact of potential errors in evaluation for traits with different genetic control. Collectively, we tested three analysis procedures (CONT, CAT-CONT, and CAT), evaluated under three model approaches (Mixed models, Bayesian, and Random Forest), for four category levels (1–2, 1–3, 1–5, and 1–9), with three different noise levels (low, moderate and high), and two hypothetical genetic architectures (qualitative and quantitative). These multiple scenarios create high complexity to discuss the results. To circumvent this, we structured our discussion in the following format. First, we framed our narrative in terms of genomic prediction, and discussed the importance of using different categorical levels to classify continuous traits. Sequentially, we emphasized the impact of using different methods (LMM and GLMM, BLRM and BORM, RFR and RFC) and distributions (continuous and categorical) on predictive accuracy. After discussing the importance of our results for prediction, we focused on inference and considered potential impacts on estimating marker effects, heritability, and genetic gains. Finally, we applied our findings of the simulated populations to a real population of blueberry.

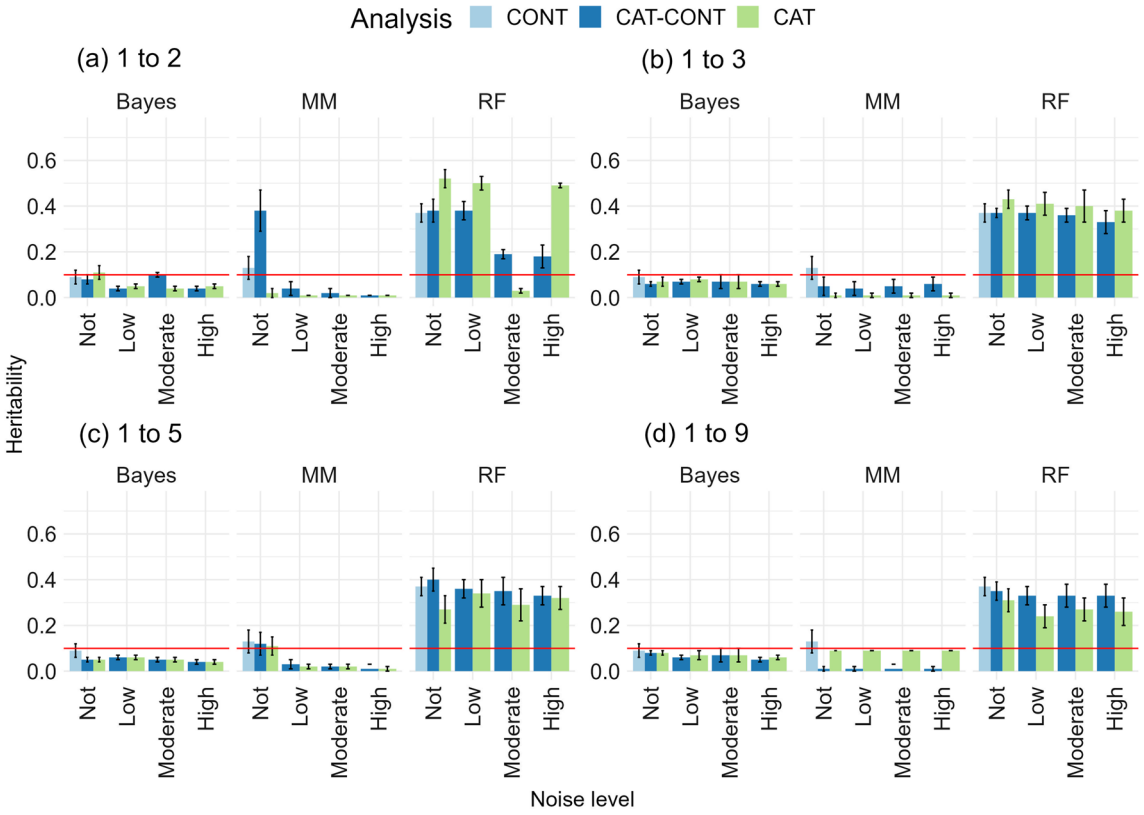

Fig. 5 Heritability estimated by different methods (Bayesian Ordinal and Bayesian Linear Regression Models (Bayes), Generalized and Linear Mixed Models (MM), and Random Forests Regression and Classification (RF)) for simulated categorical traits, with different category levels (a—1 to 2; b—1 to 3; c—1 to 5; d—1 to 9), under different levels of noise (low—20% misclassification, moderate—50% misclassification, and high—70% misclassification) and no errors, relative to continuous traits. The traits had a quantitative genetic architecture, with 100 QTLs and heritability equal to 0.10. The red line represents the simulated heritability value

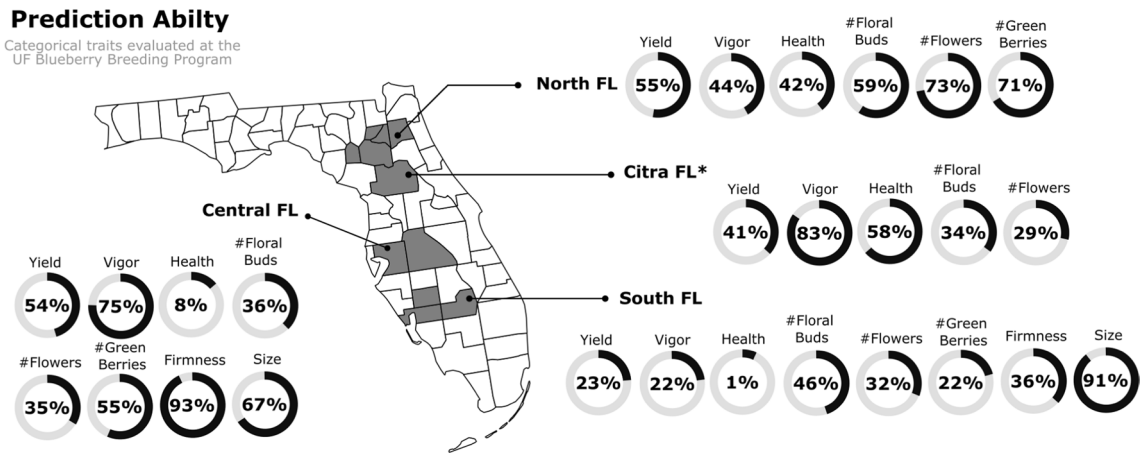

Fig. 6 Genomic prediction was performed using cross-validation procedures. The mean predictive ability using blueberry categorical phenotypes collected in several seasons (2014, 2015, 2020–2022) and breeding stages (II, III, and IV) over four macro-regions in Florida State. All predictive abilities were expressed as percentage values

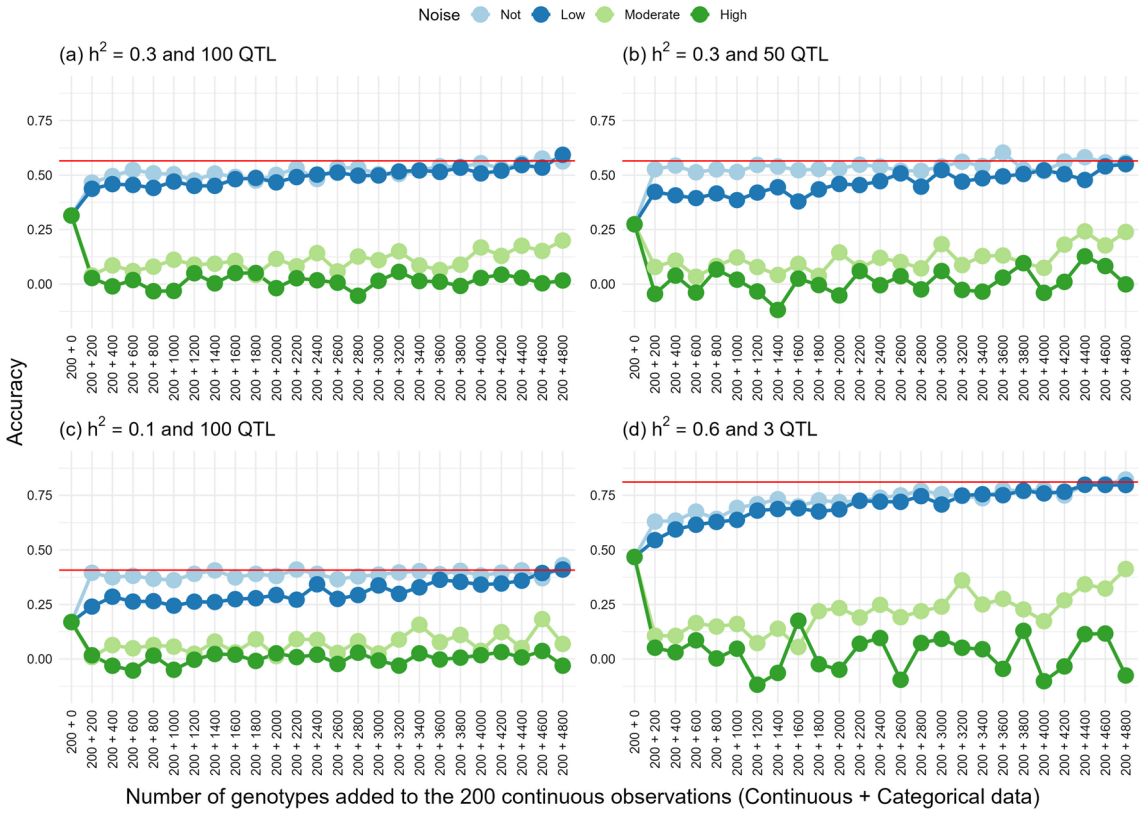

Fig. 7 Genomic prediction was performed using cross-validation procedures for simulated categorical traits, considering categories ranging from 1 to 5, different levels of noise (low–20% of misclassification, moderate–50% of misclassification, and high – 70% of misclassification), and no errors. The genetic architectures of the traits were: **a** $h^2 = 0.30$ and 100 QTLs, **b** $h^2 = 0.30$ and 50 QTLs, **c** $h^2 = 0.10$ and 100 QTLs, and **d** $h^2 = 0.60$ and 3 QTLs. The accuracies were calculated using different numbers of genotypes (ranging from 0 to 4,800 in increments of 200) recorded in the categorical trait, in addition to the 200 genotypes recorded in the continuous distribution. The horizontal red line represents the accuracy of 5,000 genotypes recorded in the continuous distribution in each scenario

What is the impact on predicting categorized traits that are continuous by nature?

Our first contribution aimed to investigate the impact of categorizing traits, that are continuous by nature. To this end, we tested different levels of categories, mimicking visual scales that have been commonly reported in the literature. For example, categories ranging from 1 to 5 are frequently used for yield evaluation (Williams et al. 2021). Disease progress, on the other hand, is often recorded either using binary evaluations (Manichaikul and Broman 2009) or 1–9 visual scales (Ferrão et al. 2019). After testing the impact of using different categorical levels and errors of evaluations, we noticed that using binary evaluations and 1–9 scores tended to result in poor results. While the use of binary evaluations tends to be a simplistic approach, a valid argument for increasing the number of categories is that results could resemble a continuous distribution that ultimately makes the use of traditional LMM models more appealing. However, when increasing the number of categories, we are also leveraging potential misclassification, in particular for intermediate evaluations–since scoring extreme results, such as high and low production, led to less subjectivity than assessing intermediate scores. The higher rates of misclassifications are probably affecting the use of 1–9 scores, making this category less efficient. Overall, we observed that using 1–3 and 1–5 scales for categorizing continuous data showed a good compromise between accuracy and error.

LMM and BLRM are robust, but not the best predictive performance for categorical traits

Our second contribution to this study relies on investigating a diverse set of statistical modeling. When the prediction is framed for categorical data, the use of GLMM and BORM are more flexible and have a broader application. However, it has the cost of being more computationally demanding, and to require a higher level of theoretical understanding to be applicable, when contrasted to standard methods (LMM and BLRM).

As an important result, we first showed that for all traits and scenarios, the use of traditional LMM ensured robust

predictive results. Villemereuil et al. (2016) also reported that the scale on which estimation is performed in generalized models satisfies the assumptions of the breeder's equation, making it useful for expressing selection of non-Gaussian traits on this scale. Similar results were also reported by Silveira et al. (2019), after evaluating rust disease resistance in *Eucalyptus urophylla* using LMM and GLMM. Similarly, Ornella et al. (2012) also compared parametric and non-parametric models and reported that LMM had superior prediction performance. Higher prediction accuracies for the LMM method were also discussed by Heuer et al. (2016), when contrasted with GLMM. Thus, as a main message, our results indicate a certain legitimacy for breeders and biometricians who are less familiar with GLMM theory and have opted to use LMM in their routines–regardless of the data distribution.

Despite the simplicity of using LMM for data Analyzes, as an important piece of relevant information, we also showed that gains can be obtained when parametric and non-parametric methods (BORM and RFC) are used for evaluating visual scores. Although those gains are marginal when compared to LMM results, whether computational resources and runtime are not limiting factors, the use of such models that do not assume normality on the residuals can result in better results. When comparing methods, we also noticed an overlap from most of the confidence intervals for accuracy, with a clear exception for the 1–9 categories. In short, it indicates similarity between methods, that aligns with previous studies in the literature (de los Campos et al. 2013; Gianola 2013). However, it is noteworthy that the RFC exhibited greater stability across different levels of noise. This indicates that RFC is a robust method that is less sensitive to variations in the presence of noise. Consequently, if there is uncertainty or lack of knowledge regarding the level of error in the dataset, the RFC method may be a suitable choice.

Genetic parameters are better assessed using Bayesian ordinal regression models

In fact, breeding programs have relied on estimating genetic parameters to guide decisions. For example, breeders have guided their decisions based on the level of genetic control (i.e.,,, heritability), the magnitude of gene action effects, the correlations between traits, and the dynamics of genotype-by-environment interactions. At the molecular level, understanding the genetic architecture of a trait requires the estimation of the number, position, and effect size of molecular markers associated with putative QTL.

Among the key genetic parameters, recovering heritability values is an alternative for discriminating methods (Azevedo et al. (2015). Herein, as an important trend, we noticed that Bayesian ordinal regression (CAT) presented robust results in recovering the simulated values. Similar results were also reported by Kizilkaya et al. (2014), when using the Bayes Cπ linear model. Tiezzi and Maltecca (2015) evaluated the impact of computing genetic parameters using LMM and GLMM in a Bayesian framework, and also reported that GLMM captured larger proportions of the genetic variance, resulting in higher heritability values. When comparing parametric and non-parametric methods, the use of RF showed some disadvantages. For example, the covariances between predictors and variance components do not have a close form and can only be estimated through recursive expressions, such as the variance of predicted values (Chen and Zhang 2013). Consequently, if RF produces inaccurate estimates of genetic values, it will naturally impact genetic parameters.

Another important parameter for implementing marker-assisted selection is the estimated marker effects. Genetic values and putative genes can be estimated using marker effects in genomic selection and genome-wide association (GWAS) studies, respectively. The question to be considered is as follows: are the different statistical modeling approaches assessing similar information at the genomic level? When compared to the parametric simulated value, the results highlighted the robustness of the LMM Analyzes. Overall, we observed high correlations between estimated markers effects and genomic breeding values, with a large percentage of agreement when genotypes were ranked.

Genomic prediction and inference on the genetic basis of blueberry traits using categorical data

Our final contribution relies on GS implementation using real data. Over the years, blueberry breeders often faced the dilemma of phenotyping a large breeding population using visual scores or focusing on a restricted number of samples and collecting more accurate continuous phenotypes. In this context, yield is a prime example. While blueberry bushes need to be harvested multiple times over the season, the phenotyping process is labor intensive, costly, and has low throughput. To circumvent this, the UF blueberry breeding program has the following strategy: berries from mature plants on advantaged breeding stages are manually harvested and weighed, and plants from earlier stages are visually scored based on general yield, number of flowers, flower buds, and green fruits. For genomic prediction, yield collected at the continuous scale is formally used to train our predictive models, while the visual scores are used as auxiliary traits.

To effectively test the importance of using categorical data in our breeding pipeline, we selected a diverse set of traits measured at the categorical scale to be predicted using genome-base methods. To our knowledge, there are no studies in fruit crops addressing predictive performances in such diverse set of traits evaluated in large breeding populations,

in different environments. Initially, we noticed low-to-moderate predictive abilities for all the blueberry traits. As a form to validate our results, we contrasted categorical vs. continuous results using previous studies. Remarkably, the use of the Bayesian approach resulted in more reasonable values, with estimates comparable with values reported by Cellon et al. (2018) and de Bem Oliveira et al. (2020), but using firmness and size traits collected in the continuous scale. For yield, we noticed that combining a few continuous data points with a massive number of categorical information has the potential to leverage predictive Analyzes by increasing the population sizes. As the main objective of this study is to translate genomics information into breeding decisions, while the use of automatic phenotype acquisition is not in the routine of many breeding programs, we argue that combining continuous and categorical data for prediction is a valid strategy. Reducing the subjectivity of visual evaluations and screening large diversity of genotypes are important elements for practical implementation.

Conclusion

Altogether, the real and simulated data used in this investigation allow us to provide a blueprint for how visual scores could be used by plant breeders. Regarding our main research questions, we can first conclude that categorical traits can be effectively used for prediction and inference on traits with different genetic architecture, with gains and precision directly related to the amount of noise and subjectivity included in the Analyzes. Secondly, the use of traditional statistical approaches showed robustness, but not the best predictive results over different error levels and genetic architectures. Thus, when time and computational resources are not a barrier, the use of Bayesian ordinal regression models are preferable. Next, we reported large predictive abilities for a group of categorical traits collected in blueberry, which opens important venues to include such traits in a molecular breeding pipeline. Finally, we integrated continuous and categorical data and simulated scenarios of genomic prediction for traits with different genetic architectures. Simply stated, we suggested that by including 600–1000 categorical data phenotypes with low error, we can verify the improvement of stable predictive performance. At this point, breeders are encouraged to reflect on the importance of allocating resources to training their team and the costs related to phenotyping. In the case of blueberry, it is noteworthy that collecting yield and other fruit set traits over the seasons is costly and time-consuming. Investing in better training associated with visual scores is a more feasible alternative when phenotyping is designed for larger populations–at least until the use of computer vision methods for high-throughput phenotyping is fully adopted.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-023-04512-w.

Acknowledgements This work was supported by the UF royalty fund generated by the licensing of blueberry cultivars. CFA, MDVR and MN are grateful for the CNPq research fellowship.

Author contribution statement The authors confirm contribution to the paper as follows: PM, LFVF and CFA conceived and supervised the study. JB coordinated the collection and genotyping of the samples. LFVF, CFA, MDVR, MN and ACCN analyzed and interpreted the genomic selection results. LFVF and CFA wrote the paper and included the revision from all authors. All authors read, reviewed and approved the final version of the manuscript for publication.

Funding This work was supported by the UF royalty fund generated by the licensing of blueberry cultivars.

Data availability The data supporting the conclusions of this article will be shared by open request to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Azevedo CF, de Resende MDV, Silva FF et al (2015) Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genet 16:105. https://doi.org/10.1186/s12863-015-0264-2

Batista LG, Gaynor RC, Margarido GRA et al (2021) Long-term comparison between index selection and optimal independent culling in plant breeding programs with genomic prediction. PLoS ONE 16:e0235554. https://doi.org/10.1371/journal.pone.0235554

Benevenuto J, Ferrão LF, Amadeu RR, Munoz FP (2019) How can a high-quality genome assembly help plant breeders? Gigascience. https://doi.org/10.1093/gigascience/giz068

Biffani S, Pausch H, Schwarzenbacher H, Biscarini F (2017) The effect of mislabeled phenotypic status on the identification of mutation-carriers from SNP genotypes in dairy cattle. BMC Res Notes 10:230. https://doi.org/10.1186/s13104-017-2540-x

Butler D (2022) asreml: fits the linear mixed model. In: R package version 4.1.0.160

Cellon C, Amadeu RR, Olmstead JW et al (2018) Estimation of genetic parameters and prediction of breeding values in an autotetraploid blueberry breeding population with extensive pedigree data. Euphytica 214:87. https://doi.org/10.1007/s10681-018-2165-8

Chen Z, Zhang W (2013) Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. PLoS Comput Biol 9:e1002956. https://doi.org/10.1371/journal.pcbi.1002956

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Res 19:136–142. https://doi.org/10.1101/gr.083634.108

Colle M, Leisner CP, Wai CM et al (2019) Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. Gigascience. https://doi.org/10.1093/gigascience/giz012

Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Bem OI, Amadeu RR, Ferrão LFV, Muñoz PR (2020) Optimizing whole-genomic prediction for autotetraploid blueberry breeding. Heredity (edinb) 125:437–448. https://doi.org/10.1038/s41437-020-00357-x

de Campos G, Hickey JM, Pong-Wong R et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. https://doi.org/10.1534/genetics.112.143313

Ferrão LFV, Ferrão RG, Ferrão MAG et al (2019) Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. Heredity (edinb) 122:261–275. https://doi.org/10.1038/s41437-018-0105-y

Ferrão LF, Amadeu RR, Benevenuto J et al (2021) Genomic selection in an outcrossing autotetraploid fruit crop: lessons from blueberry breeding. Front Plant Sci. https://doi.org/10.3389/fpls.2021.676326

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing

Gaynor RC, Gorjanc G, Hickey JM (2021) AlphaSimR: an R package for breeding program simulations. G3 Genes Genomes Genet. https://doi.org/10.1093/g3journal/jkaa01

Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194:573–596. https://doi.org/10.1534/genetics.113.151753

Gilmour AR, Anderson RD, Rae AL (1985) The analysis of binomial data by a generalized linear mixed model. Biometrika 72:593. https://doi.org/10.2307/2336731

González-Recio O, Rosa GJM, Gianola D (2014) Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest Sci 166:217–231. https://doi.org/10.1016/j.livsci.2014.05.036

Harville DA, Mee RW (1984) A mixed-model procedure for analyzing ordered categorical data. Biometrics 40:393. https://doi.org/10.2307/2531393

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York

Heuer C, Scheel C, Tetens J et al (2016) Genomic prediction of unordered categorical traits: an application to subpopulation assignment in German Warmblood horses. Genet Sel Evol 48:13. https://doi.org/10.1186/s12711-016-0192-2

Ho TK (1995) Random decision forest. In: 3rd international conference on document analysis and recognition. Montreal, pp 278–282

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York

Kizilkaya K, Fernando RL, Garrick DJ (2014) Reduction in accuracy of genomic prediction for ordered categorical data compared to continuous observations. Genet Sel Evol 46:37. https://doi.org/10.1186/1297-9686-46-37

Liaw A, Wiener M (2022) Classification and regression by randomForest. R news 2(3):18–22

Manichaikul A, Broman KW (2009) Binary trait mapping in experimental crosses with selective genotyping. Genetics 182:863–874. https://doi.org/10.1534/genetics.108.098913

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Merrick LF, Lozada DN, Chen X, Carter AH (2022) Classification and regression models for genomic selection of skewed phenotypes: a case for disease resistance in winter wheat (*Triticum aestivum* L.). Front Genet. https://doi.org/10.3389/fgene.2022.835781

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P et al (2015a) Genomic prediction models for count data. J Agric Biol Environ Stat 20:533–554. https://doi.org/10.1007/s13253-015-0223-4

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P et al (2015b) Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. G3 Genes Genomes Genet 5:291–300. https://doi.org/10.1534/g3.114.016188

Montesinos-López OA, Montesinos-López A, Crossa J (2017) Bayesian genomic-enabled prediction models for ordinal and count data. Genomic Selection for Crop Improvement. Springer, Cham, pp 55–97

Montesinos-López A, Gutierrez-Pulido H, Montesinos-López OA, Crossa J (2020a) Maximum *a posteriori* threshold genomic prediction model for ordinal traits. G Genes Genomes Genet 10:4083–4102. https://doi.org/10.1534/g3.120.401733

Montesinos-López OA, Montesinos-López JC, Singh P et al (2020b) A multivariate poisson deep learning model for genomic prediction of count data. G3 Genes Genomes Genet 10:4177–4190. https://doi.org/10.1534/g3.120.401631

Montesinos López OA, Montesinos López A, Crossa J (2022a) Multivariate statistical machine learning methods for genomic prediction. Springer, Cham

Montesinos López OA, Montesinos López A, Crossa J (2022b) Bayesian and Classical prediction models for categorical and count data. Multivariate statistical machine learning methods for genomic prediction. Springer International Publishing, Cham, pp 209–249

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135:370. https://doi.org/10.2307/2344614

Ornella L, Singh S, Perez P et al (2012) Genomic prediction of genetic values for resistance to wheat rusts. Plant Genome. https://doi.org/10.3835/plantgenome2012.07.0017

Pérez P, de Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442

Pérez-Rodríguez P, Flores-Galarza S, Vaquera-Huerta H et al (2020) Genome-based prediction of Bayesian linear and non-linear regression models for ordinal data. Plant Genome. https://doi.org/10.1002/tpg2.20021

Resende MDV de, Azevedo CF, Nascimento M, et al (2018) Modelos Hierárquicos Generalizados Lineares Mistos (HGLMM), Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP

Schielzeth H, Dingemanse NJ, Nakagawa S et al (2020) Robustness of linear mixed-effects models to violations of distributional assumptions. Methods Ecol Evol 11:1141–1152. https://doi.org/10.1111/2041-210X.13434

Silveira LS, Filho M, Azevedo CF et al (2019) Research article Bayesian models applied to genomic selection for categorical traits. Genet Mol Res. https://doi.org/10.4238/gmr18490

Stroup WW (2015) Rethinking the analysis of non-normal data in plant and soil science. Agron J 107:811–827. https://doi.org/10.2134/agronj2013.0342

Tiezzi F, Maltecca C (2015) Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. Genet Select Evol 47:24. https://doi.org/10.1186/s12711-015-0100-1

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

Villemereuil P, Schielzeth H, Nakagawa S, Morrissey M (2016) General methods for evolutionary quantitative genetic inference from generalized mixed models. Genetics 204(3):1281–1294. https://doi.org/10.1534/genetics.115.186536

Williams D, Hackett CA, Karley A et al (2021) Seeing the wood for the trees: hyperspectral imaging for high throughput QTL detection in raspberry, a perennial crop species. Fruit Res 1:1–11. https://doi.org/10.48130/FruRes-2021-0007

Yabe S, Iwata H, Jannink J-L (2018) Impact of mislabeling on genomic selection in cassava breeding. Crop Sci 58:1470–1480. https://doi.org/10.2135/cropsci2017.07.0442