Received: 11 March 2024 | Accepted: 2 May 2024

DOI: 10.1002/ppj2.20109

The Plant Phenome Journal OPEN ACCESS

ORIGINAL ARTICLE

A comparison of genomic and phenomic selection methods for yield prediction in *Coffea canephora*

Paul Adunola[1] | **Estefania Tavares Flores**[1] | **Elaine M. Riva-Souza**[2] | **Maria Amélia G. Ferrão**[2,3] | **João Felipe B. Senra**[2] | **Marcone Comério**[2] | **Marcelo C. Espindula**[2,3] | **Abraão C. Verdin Filho**[2] | **Paulo S. Volpi**[2] | **Aymbiré F. A. Fonseca**[2,3] | **Romario G. Ferrão**[3,4] | **Patricio R. Munoz**[1] | **Luis Felipe V. Ferrão**[1]

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

[2]Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural—Incaper, Vitória, Espírito Santo, Brazil

[3]Empresa Brasileira de Pesquisa Agropecuária—Embrapa Café, Brasília, Distrito Federal, Brazil

[4]Multivix Group, Vitória, Espírito Santo, Brazil

Correspondence
Luis Felipe V. Ferrão, Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Gainesville, FL, USA.
Email: lferrao@ufl.edu

Assigned to Associate Editor Michael Allen Gore.

Abstract

Genomic prediction has been proposed as the standard method to predict the genetic merit of unphenotyped individuals. Despite the promising results reported in the plant breeding literature, its routine implementation remains difficult for some crops. This is the case with *Coffea canephora*, in which costs and availability of molecular tools are major challenges for most breeding programs. To circumvent this, the use of near-infrared spectroscopy (NIR) has been recently proposed as an alternative to complement marker-assisted selection. The so-called phenomic selection relies on the reflectance spectrum to capture similarities between individuals and emerges as a valid approach for prediction. With promising results reported in multiple annual crops, we hypothesize that phenomic prediction could be a cost-efficient approach to incorporate into a practical coffee breeding program. To test it, we relied on a diverse population of *C. canephora*, evaluated for yield production, in two geographical locations over four harvest seasons. Our contributions in this paper are twofold: (i) We compared phenomic and genomic selection results, and showed large predictive abilities when NIR is used as a predictor for within and across-location predictions, and (ii) we presented a critical view of how both information sets could be combined into a contemporaneous coffee breeding program. Altogether, our results show how multi-omic information could be integrated in the same framework to leverage genetic gains in the long term.

1 | INTRODUCTION

Coffee possesses significant global importance, impacting the cultural, economic, and social aspects of our society. It is estimated that more than 3 billion cups are consumed daily. It contributes to an annual income of ~$200 billion and provides jobs for 125 million people (Bozzola et al., 2021). With a complex production chain, from seed to cup, the sustainability of the crop can be affected by different factors. Plant breeding has a pivotal role in this process. Positioned at the

base of the chain, breeders have developed coffee varieties that could meet the demand of consumers, retailers, farmers, and roasters. This includes plants that combine high yield, drink quality, and more resilience to the projected environmental changes, exacerbated by the increase in temperatures and erratic rainfalls (Davis et al., 2021).

Among multiple coffee species, production and consumption are segmented into two main categories: Arabica (*Coffea arabica* species) and Robusta/Conilon (*Coffea canephora* species). Arabica represents ~60% of the global production, and it is considered the main source of drink quality. However, it is a delicate crop that is quite susceptible to diseases, and with a narrow genetic diversity, genetic progress in the species has been limited in the past decades. Likewise, recent projections expect a reduction in Arabica production of ~80% by 2050, given the expected climate changes (Davis et al., 2021; Imbach et al., 2017). In this context, Robusta/Conilon production is gaining momentum in the coffee chain, emerging as a candidate for climate-smart cultivars (Ferrão, et al., 2023). More adapted to higher temperatures and resilient to diseases, its representation in the coffee market has increased from 25% to 40% in the past three decades. While this demand is not expected to slow down, the *C. canephora* industry has important challenges for the future, including leveraging its drink quality, uniformity of production, farmer profitability, and adaptation to new production systems (WCR, 2023). With these pressing factors, it is necessary to find new methodologies that can keep pace with conventional coffee breeding programs while also accelerating genetic gains and the development of improved cultivars.

In *C. canephora* breeding programs, the use of phenotypic evaluations associated with recurrent selection has been the standard and has served as the basis for releasing most of the available Robusta/Conilon cultivars (R. G. Ferrão et al., 2019; Leroy et al., 1993; Montagnon et al., 2003). With the advent of genomic prediction methods (Meuwissen et al., 2001), it is possible to predict the genetic merit of plants during their seedling stages using DNA information. This has the potential to fasten the breeding cycles and ultimately leverage genetic gain. In coffee, progress in using genomic-based methods has been reported for *C. canephora* (L. V. F. Ferrão et al., 2017; L. F. V. Ferrão et al., 2019; M. A. G. Ferrão et al., 2023; Adunola et al., 2023) and *C. arabica* (Carvalho et al., 2023; Fanelli Carvalho et al., 2020; Sousa et al., 2019). Despite the promising results in multiple coffee traits, its implementation is not straightforward. In coffee, accessible genome-wide markers that can be used routinely are still a major barrier for most breeding programs (Mbebi et al., 2022).

To overcome the limitations underlying the use of molecular markers, recent studies have argued in favor of phenomic selection based on near-infrared spectroscopy (NIR) information (Rincent et al., 2018). The use of NIR has a long history in the agronomy field. Originally proposed to predict target traits, NIR is well known to be a high-throughput, low-cost, and nondestructive method used to estimate the reflectance of a sample for numerous wavelengths. However, its use for predicting the genetic merit of unphenotyped individuals is relatively new. The motivation behind using wavelengths is analogous to the form that molecular markers are used in genomic selection: as a metric to capture genetic similarities between individuals and perform predictions (Rincent et al., 2018; Zhu et al., 2021). While NIR has a long history in coffee research, including the determination of geographical origin (Giraudo et al., 2019), quality classification (Barbin et al., 2014; Mutz et al., 2023), and determination of caffeine content (Ayu et al., 2020), it was never tested for the prediction of complex traits. Herein, our primary hypothesis is that NIR spectra can be used to capture the genetic covariance between individuals and, similarly to the use of the genomic best linear unbiased prediction (GBLUP) method (VanRaden, 2008), it could be implemented as a cost-efficient alternative for yield prediction in coffee. Examples of success have been recently proposed in multiple crops, including soybean (Zhu et al., 2021), wheat (Krause et al., 2019; Rincent et al., 2018), maize (Lane et al., 2020), and grapevine (Brault et al., 2022).

Considering the importance of using new methods to accelerate coffee breeding, in this study we addressed the following main question: When predicting coffee yield, is the use of phenomic selection methods an efficient alternative to complement genomic selection methods? Motivated by that, we used a representative germplasm collection of *C. canephora*, evaluated in two different locations over multiple years, with the following main objectives: (i) compare the information assessed via genomic and NIR methods, (ii) evaluate the predictive performance within and across environments using both methods, and finally (iii) investigate the impact of model choices for prediction. Altogether, our results draw attention to how phenomic selection could be integrated into a coffee breeding program to maximize genetic gains.

Core Ideas

- Coffee has significant global importance, impacting the cultural, economic, and social aspects of our society.
- For genetic improvements, the availability of affordable and accessible genome-wide markers is still a major barrier for coffee breeding programs.
- The use of near-infrared spectroscopy (NIR) for phenomics showed promising results in predicting yield.
- Phenomic and genomic selection can be integrated with an NIR framework to assist coffee breeders in their decision-making.

2 | MATERIALS AND METHODS

2.1 | Plant material and phenotypic analyses

The plant genotypes used in this study are part of the coffee breeding program at the Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper), ES, Brazil. The institute has the largest germplasm bank of *C. canephora* in Brazil. Since 1988, plant accessions from different regions and origins have been collected and maintained for ex-situ conservation. In 2016, an important expansion of this collection was performed, with all plants cloned and installed in two coffee-growing regions. It includes the experimental farms in Marilândia (FEM) with latitude of 19,407˚S and longitude of 40,539˚W, and the experimental farm in Bananal do Norte (FEB), with latitude of 20,750˚S and longitude of 41,228˚W, as described by M. A. G. Ferrão et al. (2021).

Up to now, this germplasm collection has a total of 606 accessions. On average, in each location, three to five plants of each of the accessions (genotype) were cloned and planted into a single plot at a spacing of 3.0 m × 1.20 m. For this study, we focused only on the yield evaluation performed in the FEM and FEB locations. Yield was measured in kilograms after harvesting each plot and computing the ratio between the total yield (in kg) and the number of plants per plot. For the FEM location, yield data were collected during four harvest seasons (2019, 2020, 2021, and 2022), while for the FEB, we used two harvest seasons (2020 and 2021).

In this study, the term "phenotypic" refers to the target trait (in this case, yield), while we use the term "phenomics" to refer to the endophenotyped collected via NIR. For the phenotypic analyses, we relied on spatial information to estimate the empirical best linear unbiased estimates (BLUEs) from two classes of models: across-location and within-location.

For the across-location models, information on rows and columns were recorded, and modeled jointly to the season effect, per location, using the following linear model:

$$\mathbf{y} = 1' \boldsymbol{\mu} + \mathbf{X}_g \mathbf{g} + \mathbf{X}_s \boldsymbol{\tau} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector of yield data (kg) collected across multiple harvest seasons and genotypes; $\boldsymbol{\mu}$ is a vector of overall mean; $\mathbf{g}$ is a vector of fixed genetic effects, with a design matrix of $\mathbf{X}_\mathbf{g}$; $\boldsymbol{\tau}$ is a vector of fixed effects associated with harvest year, with the design matrix $\mathbf{X}_s$, and $\boldsymbol{\epsilon}$ is the vector of random residual effects, where $\boldsymbol{\epsilon} \sim \text{MVN}(0, R\sigma_r^2)$. The residual was modeled assuming a first-order separable (separate functions for row and column) autoregressive (AR) structure, accounting for a direct product (AR1 × AR1), as described by Isik et al. (2017). The following correlation matrix was fitted for the residual term: $\boldsymbol{R} = \boldsymbol{\Sigma}_c \otimes \Sigma_r$, where the column and row local spatial correlation matrices were taken as autoregressive models of order 1 (AR1) with spatial correlation parameters row and column directions, respectively. More details on the importance of using spatial models, for analyses of individual plants, and in the genomic prediction context are discussed by Elias et al. (2018).

For the so-called within-location analyses, a similar mixed model was considered, with the BLUEs estimated per season across both locations. To this end, the season effect was removed from the aforementioned phenotypic model, and a model only accounting for the genotypic and spatial information (row and columns) was considered. Across- and within-location models were fitted using the ASReml-R package (Butler et al., 2009). The BLUEs for the genotype effects were used in the subsequent prediction analyses.

2.2 | NIR data

NIR data were obtained from coffee green beans harvested in FEB in 2022. From the entire germplasm collection, NIR data were collected from 263 accessions that showed similar maturation patterns. The choice for FEB location was primarily made because of the better postharvest practices presented in this location. Importantly, the logistics underlying postharvest practices in coffee are laborious. In such a diverse coffee population, it requires multiple steps, that include tracking the right time for harvesting genotypes with different maturation times, selective picking, drying the beans, grinding, and collecting the NIR information. To this end, from each genotype, a sample of ∼2 kg of ripened fruits was selected, picked, dried, and depulped (getting rid of the skin and mucilage) to obtain the coffee green beans. Each sample was ground until it formed a fine powder.

For NIR collection, three technical replicate reflectance data points were collected within the range of 906–1676 nm with a 6 nm step, consisting of ∼125 wavelength data points. Spectra data were obtained using the VIAVI MicroNIR OnSite-W device along with its corresponding software. To ensure data quality, we conducted checks for outliers and applied pretreatment techniques using the protocols described in the waves R package (Hershberger et al., 2021). The spectra data were normalized (centered and scaled), and their second derivative was computed using a Savitzky–Golay filter with a window size of 15 data points, as implemented in the waves R package (Hershberger et al., 2021).

From the transformed NIR data, we estimated variance components for genotypes, permanent environment, and residuals at each wavelength to check heritability and repeatability values. Briefly, we use a repeatability model that is commonly applied in animal science and fruit breeding literature (Hernandez et al., 2020) when technical replications are collected from the same individual (in this case, NIR information collected multiple times from the same genotype). The following linear mixed model was used: $\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z}\mathbf{u} + \mathbf{W}\mathbf{p} + \mathbf{e}$, here $\mathbf{y}$ is the vector of NIR data for the spectrum i; $\boldsymbol{\beta}$ is the vector of fixed effect; $\mathbf{u}$ is a vector of random

genetic effects as $\mathbf{u} \sim N(0, I\sigma^2_{\mathbf{u}}))$, where $\sigma^2_{\mathbf{u}}$ is the genetic variance component; $\mathbf{p}$ is the vector of permanent environmental effect and non-additive genetic that is independently distributed with means of zero and variance $\sigma^2_{\mathbf{pe}}$; and $\mathbf{e}$ is the vector of residual effect following a normal distribution with zero mean and (co)variance matrix σ^2_e. $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are incidence matrices relating the fixed and random effects to measurements in vector $\mathbf{y}$.

2.3 | Genotypic data

The 263 selected accessions from the germplasm collection were genotyped by the LGC/RAPiD Genomics company using the sequencing capture methodology. Sequencing was performed using an Illumina HiSeq2000 platform, considering 100-cycle paired-end runs. Raw reads were filtered by quality and trimmed, while the filtered reads were mapped against the *C. canephora* genome assembly (Denoeud et al., 2014) using the BWA v.0.7.17 software (Li & Durbin, 2009). Single-nucleotide polymorphisms (SNPs) were called using FREEBAYES v.1.0.1 (Garrison & Marth, 2012), targeting 10,000 probe regions designed for the sequence capture approach, as described by Resende (2016) and Alkimim et al. (2018). Loci were also filtered by applying the following criteria: minimum mapping quality of 30; only biallelic locus; maximum missing data of 40%; minor allele frequency of 1%; and minimum and maximum mean depth of 3 and 200, respectively. The remaining missing genotypes were imputed using the default parameters from BEAGLE (Browning & Browning, 2007). In total, 149,970 raw SNPs were originally reported. After following the quality-control steps, a total of 52,456 markers were retained.

2.4 | Phenomic selection models

To implement phenomic selection using NIR as predictors, we tested different models. To this end, we used the pretreated spectra data, and all NIR matrices were centered and scaled over the samples. Here, we grouped the phenomic selection models into three main categories: mixed models, Bayesian approaches, and machine learning.

For mixed models, genetic values were predicted using the BLUP and restricted maximum likelihood approach to estimate variance components, as follows: $\mathbf{y} = 1' \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{e}$; where $\mathbf{y}$ is the vector of the measured phenotype (yield); $\boldsymbol{\mu}$ is the overall vector mean; $\mathbf{Z}$ is an incidence matrix linking observations in the vector $\mathbf{y}$ to their respective genetic values; $\mathbf{u}$ is a random genetic value vector, where $\mathbf{u} \sim \text{MVN}(0, \mathbf{K}\sigma^2_k)$. $\mathbf{K}$ is the phenomic relationship matrix constructed from the transformed near-infrared spectra, and σ^2_k is the phenomic variance. $\mathbf{K}$ was computed as $\frac{\mathbf{SS}'}{n_w}$, where $\mathbf{S}$ is the centered and scaled matrix for each wavelength and n_w is the number of wavelengths reflectance. A mixed model was implemented in the ASReml-R (Butler et al., 2009).

For the Bayesian approaches, we tested four regression methods with different assumptions (and distributions) for the regression parameter (effect of wavelength reflectance). For Bayesian ridge regression, we assumed that predictors follow a normal distribution, described as $\beta_j|\sigma^2_\beta \sim N(0, \sigma^2_\beta)$, where σ^2_β is a common variance associated with each wave effect. The second approach is referred to in the genomic selection literature as BayesA, and the regression parameter β_j also follows a normal distribution, described as $\beta_j|\sigma^2_{\beta_j} \sim N(0, \sigma^2_{\beta_j})$, where $\sigma^2_{\beta_j}$ represents the variance associated with each effect. In the so-called BayesB, the regression parameter β_j was modeled as a mixture distribution, $\beta_j|\sigma^2_{\beta_j} \sim \pi N(0, \sigma^2_{\beta_j}) + (1-\pi)\text{DG}(0)$, where π represents the proportion of non-null effects and follows a beta distribution and DG(0) is a degenerate distribution centered at zero. Finally, in the Bayesian LASSO (BL), the β_j follows a double-exponential distribution, which can be expressed as $\beta_j|\sigma^2_{\beta_j}, \lambda \sim \text{DE}(0, \frac{\sqrt{\sigma^2_{\beta_j}}}{\lambda})$, where λ^2 follows a gamma distribution. The variance terms, σ^2_β and $\sigma^2_{\beta_j}$, follow a scaled-inverse χ^2 density. More details about the Bayesian methods here reported are discussed by Pérez and de los Campos (2014). All Bayesian models were implemented in the BLGR R package (Perez & de los Campos, 2014) with a total of 120,000 iterations, including a burn-in period of 20,000 iterations and thinning every five iterations.

We also tested some methods described in machine-learning literature (James et al., 2013). We also explored nonlinear models, such as random forest (RF)—a method based on tree algorithms—and support vector machine radial (SVMR), a supervised learning algorithm that aims to find a regression hyperplane to minimize the distance from the sample point farthest from the hyperplane using a radial basis function as the kernel function. Specifically, RF was implemented using the randomForest R package (Liaw & Wiener, 2002), while the SVMR was implemented using the e1071 R package (Dimitriadou et al., 2009). The use of partial least square regression (PLSR) relies on a dimension-reduction approach. We used the pls R package (Mevik & Wehrens, 2015) for fitting the PLSR models.

2.5 | Genomic and multi-omic predictions

For genomic prediction, we relied on GBLUP analyses to compute genomic estimated breeding values (VanRaden, 2008). Briefly, we used the same mixed model described for the NIR information, but the relationship information among individuals was computed using molecular marker

information (the **G** matrix). The **G** matrix was estimated in the AGHmatrix R package (Amadeu et al., 2016, 2023). A mixed model was implemented in the ASReml-R (Butler et al., 2009).

We also expanded the mixed model analyses to incorporate phenomic and genomic information into the same framework. To this end, we used a multi-kernel model where we included one random effect associated with genomic information and a separated random effect associated with phenomic as follows: $\mathbf{y} = 1'\mu + \mathbf{Z}_g u + \mathbf{Z}_v v + \boldsymbol{e}$, where $\mathbf{y}$ is the vector of the measured phenotypes (yield); μ is the overall mean; $\mathbf{Z}_g$ and $\mathbf{Z}_v$ are incidence matrices linking observations in the vector $\mathbf{y}$ to their respective genomic and phenomic values, respectively; u and v are both independent random effects, normally distributed and with variance and covariance terms defined as kernel matrices constructed using genomic and NIR information. The **G** matrix was estimated in the AGHmatrix R package (Amadeu et al., 2016, 2023) as: $\mathbf{G} = \mathbf{M}M'/2\sum p_k(1-p_k)$, where **M** is the centered and scaled molecular marker matrix and p_k is the allele frequency (VanRaden, 2008). A multi-kernel mixed model was implemented in the ASReml-R (Butler et al., 2009). Heritability was computed as the ratio between the genetic and the residual term, with genetic components estimated via genomic models. We reported the proportion of the phenotypic variation (PVE) explained by phenomic selection models as the ratio between the phenomic and the residual term.

2.6 | Cross-validation scheme

The predictive performance was computed as the Pearson's correlation between the predicted and the adjusted mean values. Herein, we tested predictions within- and across-locations. For within-location predictions, a 10-fold cross-validation scheme repeated five times was used. For validation across-locations, in each site (FEB and FEM), we first corrected the yield values accounting for the year and spatial variations. With the empirical BLUEs computed; we calibrated phenomic and genomic models in one location and predicted the other.

3 | RESULTS

3.1 | Phenotypic variation

The germplasm panel underlying this study consisted of 606 *C. canephora* accessions evaluated in two locations (FEB and FEM) in different harvest seasons. For yield production, we could first notice different patterns of phenotypic correlations across locations and years (Figure 1a). Within-locations, low to moderate correlation values were reported between 2020 and 2021 (0.32 in FEB) and 2021 and 2022 (0.31 in FEM). Interestingly, field data collected in 2019 in FEM were more similar to the patterns observed in the FEB location than for the other years collected in FEM. Mean yield values were also highly different across the environments, with the highest value reported for FEB_2021 (on average, 5.1 kg per plant) and the lowest for FEM_2020 (on average, 1.1 kg per plant). Altogether, these results shed light on the relevance of spatial and temporal variation in coffee (i.e., genotype-by-environment interaction). The importance of genotype-by-environment interactions was already discussed by other studies that indicated different genetic variances and ranking changes for genotypes evaluated in this same macro-region (Adunola et al., 2023; L. F. V. Ferrão et al., 2017; L. F. V. Ferrão, et al., 2019). Adjusted mean values (BLUEs) were slightly higher for FEB when compared to the FEM location (Figure 1b).

3.2 | NIR information

From the green beans, multiple NIR measurements were collected individually per genotype. The transformed NIR spectra covered the range from 900 to 1700 nm and included 125 wavelengths (Figure 2a). We carried out correlation analyses between the wavelength spectrum and noticed small correlation clusters (Figure 2b), with moderate Pearson's correlation values (Figure 2b). This ultimately affected the variance components, which showed a similar projection across all the predictors (Figure 2c). Moderate-to-high heritability values were computed for all wavelengths, with a low variation reported for the permanent environment effect, which indicates good repeatability between the technical repetitions collected in this study. We finally correlated each wavelength and the BLUEs estimated in the FEM and FEB locations (Figure 2d and Figure S1). The results reveal correlation coefficients ranging from −0.2 to 0.2, thus without a single predictor explaining a large portion of the phenotypic correlation. While fairly small, these correlations were consistent across locations, underlining some stability across locations of the relationship between yield and NIR.

3.3 | Genomic information

Molecular markers are spanning the entire genome with very few gaps (Figure 3a). Using NIR and genomic information, we analyzed the population structure in the coffee germplasm (Figure 3b). Using genome-wide marker data, we clearly separated the accessions into three main groups. These clusters refer to the two botanical groups ("Robusta" and "Conilon") presented in *C. canephora* species; while genotypes classified as "Hybrids" are derived from crosses between both

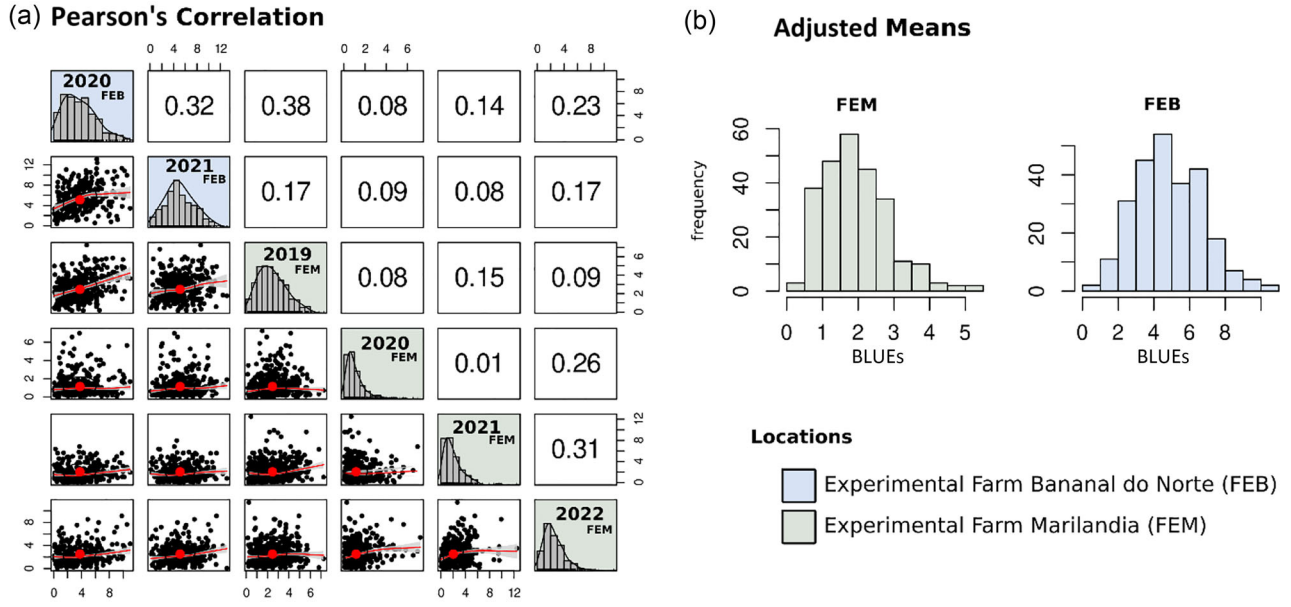

FIGURE 1 (a) The Pearson's correlation for yield prediction measured in two different locations (experimental farm in Bananal do Norte [FEB] and experimental farms in Marilândia [FEM]) across multiple harvest seasons (2019, 2020, 2021, and 2022). (b) Distribution of the empirical best linear unbiased estimators (BLUEs) for each location (FEB and FEM) after correcting for harvest and spatial effects.

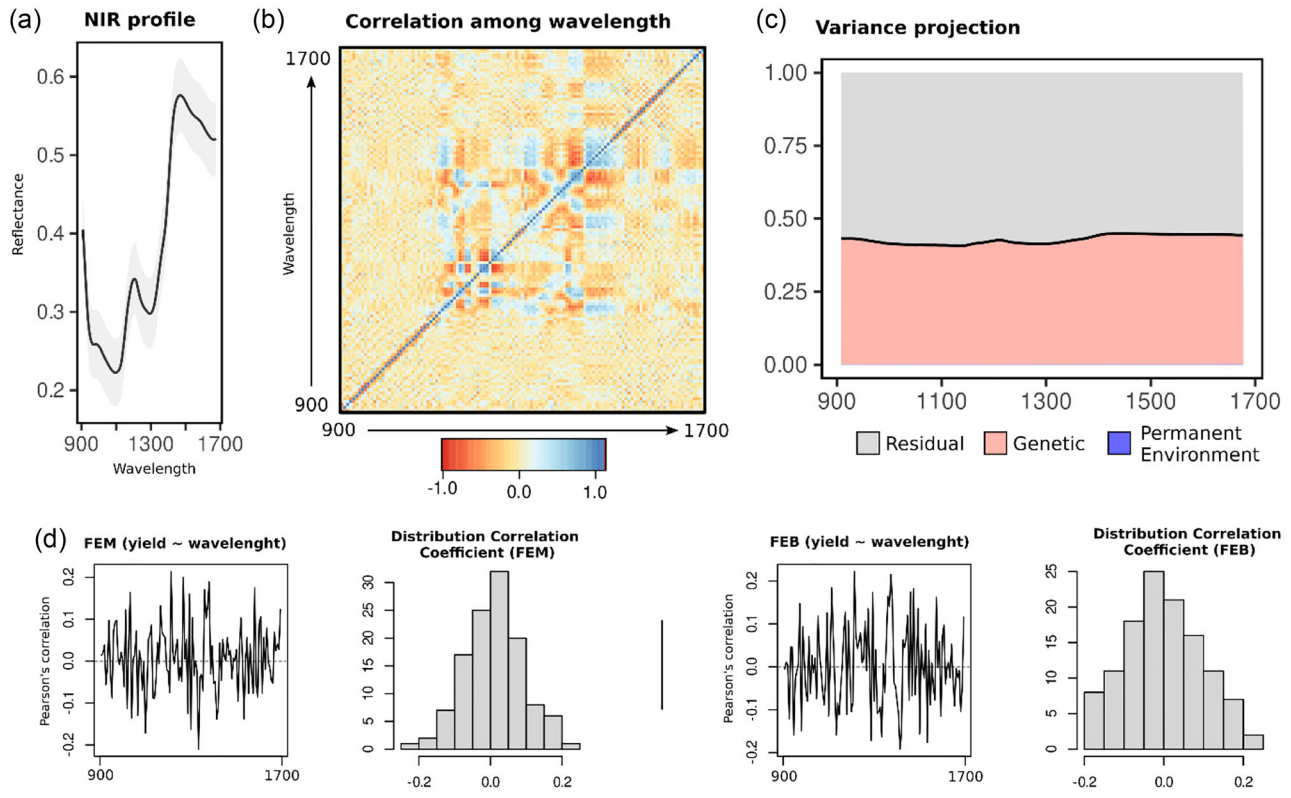

FIGURE 2 (a) Original near-infrared spectroscopy (NIR) profile of coffee green beans; the black line shows the average value, and the gray color represents the standard deviation of the population of individuals. (b) Correlation among 125 pre-treated wavelengths collected in coffee green beans. (c) Proportion of the phenotypic variance explained by the genetic term, permanent environment (as a metric of repeatability of the trait), and residual effects computer per pre-treated NIR spectrum. The blue part (permanent environment) is minimal (i.e., low values). (d) The Pearson's correlation (and their distribution) between coffee yield and pre-treated NIR spectrum over the two different locations (experimental farms in Marilândia [FEM] and experimental farm in Bananal do Norte [FEB]).

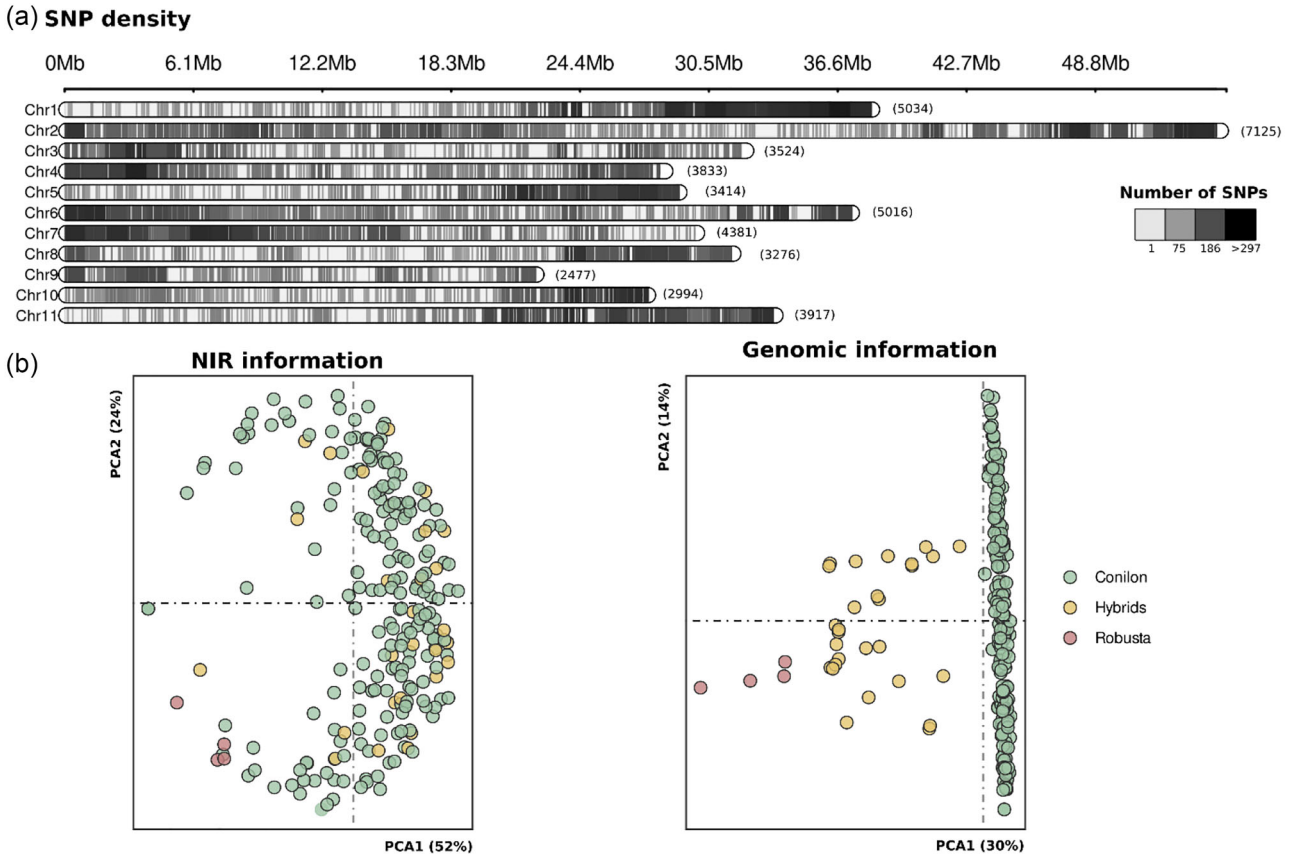

FIGURE 3 (a) Distribution of single-nucleotide polymorphisms (SNPs) across the 11 chromosomes of *Coffea canephora*. Numbers close to each chromosome indicate the total number of SNPs. (b) Principal component analyses (PCA) of 254 coffee accessions evaluated via near-infrared spectroscopy (NIR) and genome-wide marker information. Both PCAs were created after using NIR and SNP information to compute the relationship between individuals in the germplasm collection.

groups. Importantly, genomic classification matches the phenotypic and botanical coffee descriptors, already reported in this germplasm collection (M. A. G. Ferrão et al., 2021). In contrast, the NIR information was not able to detect this expected population structure. Even with the first principal components explaining more of the variation, only one cloud with a large overlap between the three groups was reported.

3.4 | Prediction models for phenomic selection

For phenomic selection, 125 NIR wavelengths were used as independent variables to predict coffee yield. Before testing it for yield prediction within and across environments, we first tested different statistical methods to predict the BLUEs computed in the FEM location. The choice for predicting the FEM location was motivated by (i) the larger number of harvest seasons reported in this site and (ii) to reduce potential prediction biases created by within-location prediction, since the NIR data were collected in the FEB location. To test it, we fitted eight models, including linear and nonlinear approaches, in their mixed model, machine learning, and Bayesian versions (Table 1).

Overall, models presented a similar predictive performance with accuracy ranging from 0.32 to 0.38. The only exception was the SVM linear method, which presented a slightly lower predictive performance (0.28). PLSR showed the best predictive abilities, followed by the SVMR model (0.38 and 0.36, respectively). The use of Bayesian models, with different assumptions, did not result in better predictive accuracies. Traditional mixed models (BLUPs) showed reasonable results (0.32) when compared to other methods that are more computationally intensive. We then used BLUP analyses for genomic and phenomic prediction.

3.5 | Comparing genomic and phenomic models for within and across-environment predictions

Prior to any predictive analyses, we quantified the heritability and PVE values associated with the yield trait per environment (Figure 4a). Using phenomics and genomics, extremely

TABLE 1 Comparing eight phenomic models to predict coffee yield using near-infrared spectroscopy (NIR). Prediction accuracy was measured as the Pearson's correlation between predicted and best linear unbiased estimator (BLUEs) computed in the experimental farms in Marilândia (FEM) location after corrected by spatial and harvest season effects. The prediction error was measured using the root square mean square error (RMSE).

Method	Assumption	Class	Accuracy	RMSE
BLUP	Linear	Mixed model	0.32	2.99
BayesA	Linear	Bayesian	0.33	2.12
BayesB	Linear	Bayesian	0.32	0.88
LASSO	Linear	Bayesian	0.32	0.89
PLSR	Linear	Linear model	0.38	0.89
Random forest	Nonlinear	Machine learning	0.31	0.91
SVM linear	Linear	Machine learning	0.28	1.02
Support vector machine radial	Nonlinear	Machine learning	0.36	0.87

Abbreviations: BLUP, best linear unbiased prediction; LASSO, least absolute shrinkage and selection operator; PLSR, partial least square regression; SVM, support vector machine.

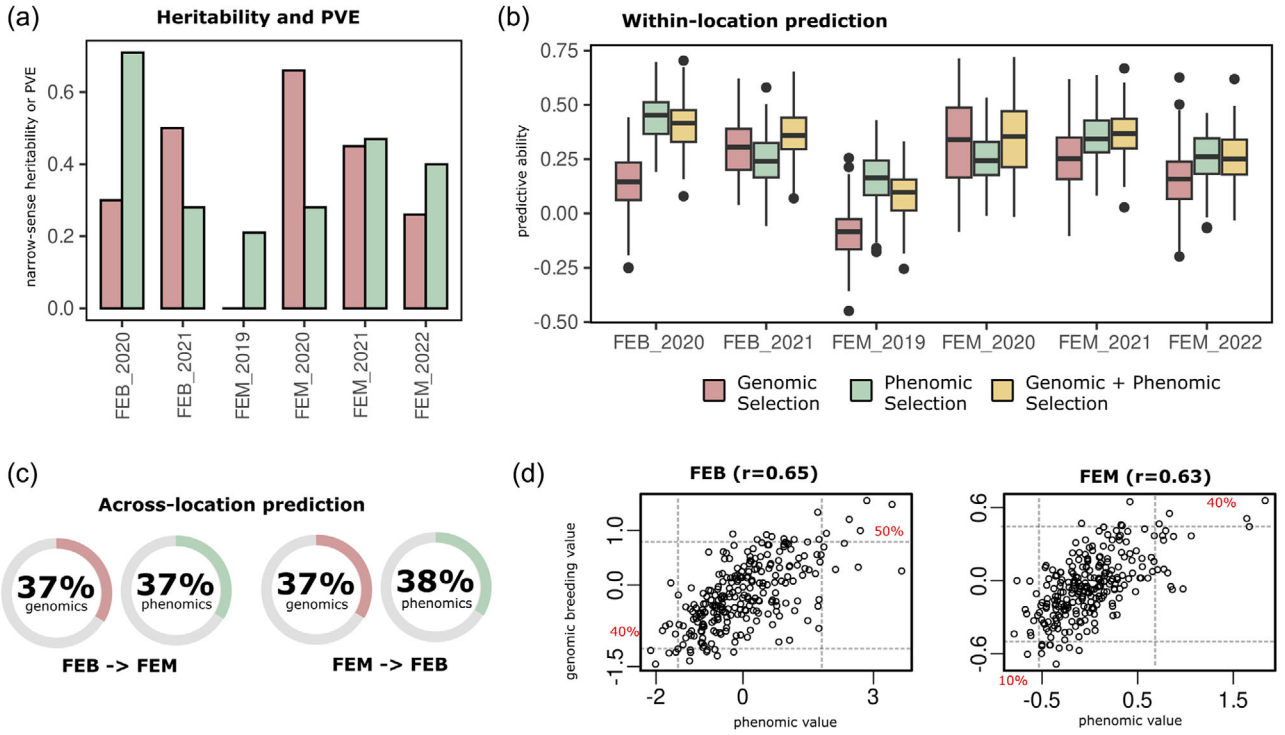

FIGURE 4 (a) Genomic heritability (in red) computed across different locations (experimental farms in Bananal do Norte [FEB] and Marilandia [FEM]) and different harvest seasons (2019, 2020, 2021, and 2022). Values were compared to the percentage of variance explained (PVE) computed using NIR information (in green). (b) Predictive ability computed in a *k*-fold cross-validation scheme, as the Pearson's correlation between phenotypes and genomic/phenomic prediction across different environments. (c) Predictive ability was measured in percentage, when genomic and phenomic models were trained in a location to predict other. (d) The Pearson's correlation between breeding values computed via phenomics and genomics in two locations (FEB and FEM). Values in red indicate the agreement between both methods when selecting the best and worst 10 genotypes in the population.

low values were reported for the 2019 season, which is mostly associated with the data quality collected in a year of extreme drought. At the genomic level, low to moderate heritability values were reported across the environments, a fact that was expected given the complex genetic architecture associated with yield data. For the NIR data, larger heritability values were reported in the FEB location, where the spectrum data were collected.

For prediction analyses, the same set of individuals was used for genomic and phenomic analyses. Thus, we first investigated within-location prediction. Results mirrored the heritability values, and, in general, the use of phenomic analyses

showed better results than genomic selection (Figure 4b). When phenomics and genomic information were modeled together, higher prediction accuracies were reported in multiple scenarios (FEB_2021, FEM_2020, FEM_2021, and FEM_2022). For across-location predictions, genomic and phenomic selections showed good (and similar) predictive performances (Figure 4c).

Finally, we used the BLUEs estimated in each location (FEB and FEM) to compute the genomic breeding values and the phenomic values (Figure 4c). We compared both results using the Pearson's correlation. Overall, both vectors showed positive and large correlation values (0.65 and 0.63 for FEB and FEM locations, respectively). As part of the coffee breeding program, breeders usually care about selecting the best individuals to be either used as parents or cloned. Moderate agreement between selections performed via genomic and NIR information was reported in both locations when the top 10 individuals were selected per environment (40% and 50% of agreement in the FEB and FEM locations, respectively).

4 | DISCUSSION

Genomic selection has become a routine tool to assist selection in plant breeding (Hickey et al., 2017). While promising predictive abilities are reported in multiple crops, its practical implementation in coffee has encountered some challenges. With most of the coffee breeding programs supported by the public sector in developing countries, the costs and logistics associated with high-throughput genotyping are still key barriers faced by breeders (Mbebi et al., 2022). In this research, we sought that phenomic selection, guided by NIR information, could be an alternative to genomic analyses to predict yield performance in coffee.

4.1 | Why NIR for phenomic selection in coffee?

Coffee is a perennial species with a long juvenile period that requires large areas for cultivation. With selections primarily guided by visual and phenotypic assessment, genetic improvements were revealed to be costly, slow, and labor-intensive (Ferrão,et al., 2023). The use of phenomic selection emerges as an alternative to make this process less subjective, high throughput, and more cost-efficient in the long term. In a recent study, Mbebi et al. (2022) trained a phenomic model using chlorophyll fluorescence (ChlF) to predict growth-related traits in *C. arabica*. The authors reported better prediction abilities than genomic selection for multiple traits, a fact that opens new alternatives for coffee breeders.

Herein, for the first time in coffee literature, we relied on the use of NIR information for predicting yield in *C. canephora*. Mainly motivated by a large number of studies reported in cereals and grains (Cuevas et al., 2019; Lane et al., 2020; Rincent et al., 2018; Zhu et al., 2021), the motivation behind our study was to focus on spectra results to build relationships between individuals and capture some Mendelian sampling information. Simply stated, it would work analogously with molecular markers in the GBLUP context, where we expect higher predictive performance than using relationship information derived from recorded pedigree or mass selection.

From a biological perspective, the underlying question revolves around what exactly is being measured by NIR. In the literature, there are some lines of evidence agreeing on the concept of endophenotypes. Rincent et al. (2018), for example, defined endophenotype capture by NIR as "different molecular layers between the genome and the phenotype, which permit the integration of interactions and regulatory networks." In a recent publication, Zhou et al. (2013) argued that the biological principle underlying such endophenotypes is not totally clear. The central benefit is that, for practical purposes, this question becomes secondary. As far as NIR can capture relationships and outputs accurate prediction values, its use in breeding programs can be justified to aid selection.

For coffee, there are multiple lines of evidence connecting NIR and chemical information that can shed some light on the nature of the endophenotypes. With a long history in the coffee literature, NIR has been used as a form of measuring coffee quality, defining geographical origin, and as a proxy for measuring some complex traits (Ayu et al., 2020; Giraudo et al., 2019; Mutz et al., 2023). In a food science review, Barbin et al. (2014) have reported the chemical assignments for some of the most important NIR bands. Herein, we can speculate that NIR is mostly capturing differences in alcohol and hydrochloride contents among green beans while also quantifying some levels of lipids and carbohydrates. This information could partially connect the biological nature between endophenotypes and the genetic similarity computed in this study.

4.2 | Model comparison for phenomic prediction: On the relevance of BLUP analyses

When implementing predictive analyses, an important question is what statistical method might better predict unobserved environments (L. F. V. Ferrão et al., 2019). With a large number of methods proposed in the literature, in general, all approaches shared the same underlying purpose: it includes handling highly dimensional data, where the number of

variables is larger than the number of observations (de los Campos et al., 2009).

For phenomic selection in coffee, we used the same modeling philosophy described in the genomic selection literature. In total, we tested eight methods to predict empirical BLUEs computed in a single environment (FEM location). Most methods showed good predictive abilities (>0.30). PLSR was the best with predictive abilities 10% higher than traditional mixed models. Overall, such a difference in magnitude is not common in genomic selection studies. Most genomic results have reported similar predictive accuracies across models for different traits. In coffee, for example, L. F. V. Ferrão et al. (2019) reported comparable accuracy for 12 predictive models, tested for three traits, and evaluated across multiple locations and populations. In sharp contrast, the use of phenomic models ha presented more variable results when models are compared. A similar 10% difference in magnitude was recently reported in soybeans when RF and ridge regression methods were compared (Zhu et al., 2021). Using ChlF as predictors, in *C. arabica*, an impressive difference of 30% in magnitude was noticed when Bayesian and machine-learning methods were compared for family prediction (Mbebi et al., 2022). Altogether, such examples suggest a higher dependency between statistical approach, trait under evaluation, and environmental conditions.

In this study, we argued in favor of the BLUP-based models. There are several advantages to this approach. First, multiple phenomic selection studies have relied on this framework, reporting stable results across different scenarios and traits (Cuevas et al., 2019; Galán et al., 2020; Krause et al., 2019; Lane et al., 2020; Rincent et al., 2018). Herein, for all within-location predictions, we also noticed positive and large predictive abilities. At the implementation level, the use of mixed models is also convenient. In addition to a consolidated theory in the plant and animal breeding community, its extensions toward more complex structures (i.e., multi-kernel models, inclusion of genotype-by-environment interaction, and multi-trait analyses) are straightforward.

4.3 | High predictive accuracy associated with phenomic prediction

Our next contribution to this study was to compare genomic and phenomic prediction models. We first tested both approaches for predicting six environments. Remarkably, the use of NIR as predictors showed better results than genomics in four environments. For predictions across locations, we also noticed large prediction values (0.38) that suggest good performance when training models in a location to predict others, either using genomics or phenomics models. Using whole-genomic models, similar prediction abilities using across-location predictions were reported by L. F. V. Ferrão et al. (2019). Similarly, Adunola et al. (2023) discussed that stability for yield can be predicted using molecular markers. For phenomics, relying on across-location predictions is particularly important when NIR is projected to be collected at the green beans. This is because in coffee, from fruits to beans, multiple postharvest steps are necessary. It requires a certain level of organization that includes infrastructure, costs, and labor for preparing the samples. Our results suggest that such logistics could be focused on a single location since the prediction models are working well across environments.

For phenomic prediction across environments, an important note of caution is eventual biases associated with the results. In coffee, Posada et al. (2009) have reported that NIR profiles can be strongly affected by environmental factors. More recently, Dallinger et al. (2023) have addressed this topic when comparing phenomics and genomics. The authors stressed that while genomics provides a static form for predictions, the use of endophenotypes collected in the form of NIR spectrum might be highly subjected to environmental variations. Thus, phenomic prediction could introduce a source of bias in the results by yielding better results toward the environment tested. In fact, we noticed large values for the FEB location, the site where coffee green beans were originally harvested. We tried to partially address this problem by training our models in one location to predict the other. Although we noticed good predictive ability across locations, further studies on this topic are justified. The use of more contrasting environments and multiple traits are future directions to understand the impact of phenomic prediction in coffee in the long term.

4.4 | Integrating phenomics and genomics in the same pipeline

Our final contribution is to project the use of phenomics and genomics in coffee breeding. In the past years, the use of genomic-assisted breeding has been argued as a viable alternative for the future of the coffee chain (Davis et al., 2021). In the *C. canephora* species, exploration of the heterotic effect from crosses between both botanical groups (Conilon and Robusta, from the SG1 and SG2 groups, respectively) is well reported and structured in the form of reciprocal recurrent selection design (Leroy et al., 1993; Montagnon et al., 2008). Recently, Ferrão, et al. (2023) suggested a rethinking of traditional breeding designs by integrating marker-assisted selection. However, the underlying question is how phenomics could be incorporated into this framework. Or, do we still need to use genomics for predicting unobserved phenotypes?

To answer both questions, we first relied on the results reported in this study. Phenomic prediction yielded solid results in multiple scenarios. However, the integrated use of phenomic and genomic information showed the best results in multiple scenarios. This is a similar context discussed in other field areas, including recent progress in understanding fruit flavor attributes (Ferrão, et al., 2023). The use of multi-omics information has the potential to leverage the prediction abilities of complex traits and should be considered as an alternative for future coffee research.

Replacing genomics with phenomics is a more complex task and requires more reflection. Fundamentally, it is still unknown at what level additive effects are captured via phenomic models. This makes long-term projections still elusive at this point. For yield, we reported a moderate agreement when the top 10 individuals were ranked via genomic and phenomic methods, a fact that might impact parental selection and the genetic progress. Also relevant, we discussed the biases created by the endophenotypes. The bias causes overfitting toward certain environmental conditions, which can also limit its application in the long term. With those points in mind, we envision the use of NIR for specific steps in a recurrent selection breeding program. A classical recurrent selection method is divided into two main steps: (i) population improvement, by selecting the best parents to manage the frequency of beneficial alleles over time, and (ii) product development, which consists of a series of field trials, in which potential candidates are evaluated over environments until selecting a variety that can become a cultivar. We believe that NIR could have an important role to play in product development. Therefore, after defining crosses via genomic information, a larger number of families (and siblings per family) could be established in higher density nurseries, and phenomic selection could be applied to increase the selection intensity and leverage the genetic gains. A similar line of thought was recently discussed by Zhu et al. (2021) in soybeans.

Finally, another measure of caution should be taken in the interpretation of our results, as genomic and phenomic selection were applied in different coffee tissues. To start with, for phenomic selection, we relied on NIR collected in the green beans. The choice of beans was primarily motivated by the long history of using NIR in coffee for certification and flavor/chemical prediction. However, collecting data at this level requires at least one harvest of production to be measured. This fact makes genomic selection faster since molecular markers could be obtained from leaves in the early seedling stages, at greenhouse conditions within a few months of germination. With genetic gains weighted by breeding cycle, genomic selection would be two times faster than phenomics, doubling the gains when compared to NIR evaluated in the green beans. A valid extension of our work is to test NIR collected in the leaves. Studies in wheat and poplar, for example, compared different tissues for phenomic prediction and showed similar results (Rincent et al., 2018). Herein, we justified the use of NIR on green beans because it is a common practice in the coffee industry for the analyses of quality, geographical origin, and other chemical attributes of the beans (Ayu et al., 2020; Barbin et al., 2014; Giraudo et al., 2019).

5 | CONCLUSIONS

Altogether, we have demonstrated that improvements in *C. canephora* breeding programs can be accelerated using a combination of phenomic and genomic selection. In summary, we highlight two main contributions: (i) we emphasize the large predictive ability for yield when using NIR methods; (ii) we draw attention to the use of phenomic and genomic prediction in a recurrent selection breeding program with phenomics associated with product development and genomics for parental selection. Overall, when compared to traditional methods applied in coffee breeding programs, we expect that using multi-omic methods can maximize future genetic gains and accelerate the development of new cultivars.

AUTHOR CONTRIBUTIONS

Paul R. Adunola: Conceptualization; data curation; formal analysis; methodology; validation; writing—original draft. **Estefania Taveres Flores**: Data curation; methodology. **Elaine Souza**: Data curation; investigation. **Maria Amélia Ferrão**: Data curation; investigation. **João Felipe Sera**: Data curation; investigation. **Marcone Comério**: Data curation; investigation. **Marcelo C. Espindula**: Data curation; investigation. **Abraão Verdin Filho**: Data curation; investigation. **Paulo S. Volpi**: Data curation; investigation. **Aymbire Fonseca**: Data curation; investigation. **Romario Ferrão**: Data curation; investigation. **Patricio R Munoz**: Funding acquisition; project administration; resources; writing—review and editing. **Luís Felipe Ferrão**: Conceptualization; formal analysis; investigation; methodology; project administration; supervision; visualization; writing—original draft; writing—review and editing.

ACKNOWLEDGMENTS

This work was supported by Fapes (Espírito Santo Research Foundation, Brazil), CNPq (National Council for Scientific and Technological Development, Brazil), and CAPES (Brazil). Additional support was provided by the Instituto Capixaba de Pesquisa, Assitência Técnica e Extensão Rural (Incaper), and Embrapa Cafe.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sets and codes used to fit the phenomic and genomic selection models for coffee yield are available at: https://github.com/lfelipe-ferrao/phenomic_selection

ORCID

Patricio R. Munoz https://orcid.org/0000-0001-8973-9351
Luis Felipe V. Ferrão https://orcid.org/0000-0002-9655-4838

REFERENCES

Adunola, P., Ferrão, M. A. G., Ferrão, R. G., Da Fonseca, A. F. A., Volpi, P. S., Comério, M., Verdin Filho, A. C., Munoz, P. R., & Ferrão, L. F. V. (2023). Genomic selection for genotype performance and environmental stability in *Coffea canephora*. *G3 Genes, Genomes, Genetics*, *13*, jkad062. https://doi.org/10.1093/g3journal/jkad062

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., Lopes da Silva, F., Sakiyama, N. S., & Zambolim, L. (2018). High-throughput targeted genotyping using next-generation sequencing applied in *Coffea canephora* breeding. *Euphytica*, *214*, 1–18.

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende, M. F. R., & Muñoz, P. R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. *The Plant Genome*, *9*. https://doi.org/10.3835/plantgenome2016.01.0009

Amadeu, R. R., Garcia, A. A. F., Munoz, P. R., & Ferrão, L. F. V. (2023). AGHmatrix: Genetic relationship matrices in R. *Bioinformatics*, *39*, btad445 https://doi.org/10.1093/bioinformatics/btad445

Ayu, P. C., Budiastra, I. W., & Rindang, A. (2020). NIR spectroscopy application for determination caffeine content of Arabica green bean coffee. *IOP Conference Series: Earth and Environmental Science*, *454*, 012049.

Barbin, D. F., Felicio, A. L. D. S. M., Sun, D.-W., Nixdorf, S. L., & Hirooka, E. Y. (2014). Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. *Food Research International*, *61*, 23–32. https://doi.org/10.1016/j.foodres.2014.01.005

Bozzola, M., Charles, S., Ferretti, T., Gerakari, E., Manson, H., Rosser, N., & von der Goltz, P. (2021). *The coffee guide*. International Trade Centre.

Brault, C., Lazerges, J., Doligez, A., Thomas, M., Ecarnot, M., Roumet, P., Bertrand, Y., Berger, G., Pons, T., François, P., Le Cunff, L., This, P., & Segura, V. (2022). Interest of phenomic prediction as an alternative to genomic prediction in grapevine. *Plant Methods*, *18*, Article 108. https://doi.org/10.1186/s13007-022-00940-9

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*, 1084–1097. https://doi.org/10.1086/521987

Butler, D. G., Cullis, B. R., Gilmour, A. R., & Gogel, B. J. (2009). *ASReml-R reference manual* (Version 3). Queensland Department of Primary Industries and Fisheries.

Carvalho, H. F., Ferrão, L. F. V., Galli, G., Nonato, J. V. A., Padilha, L., Maluf, M. P., Ribeiro De Resende, M. F., Fritsche-Neto, R., & Guerreiro-Filho, O. (2023). On the accuracy of threshold genomic prediction models for leaf miner and leaf rust resistance in arabica coffee. *Tree Genetics & Genomes*, *19*, Article 11. https://doi.org/10.1007/s11295-022-01581-8

Cuevas, J., Montesinos-López, O., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., Burgueño, J., Montesinos-López, A., & Crossa, J. (2019). Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3 Genes, Genomes, Genetics*, *9*, 2913–2924. https://doi.org/10.1534/g3.119.400493

Dallinger, H. G., Löschenberger, F., Bistrich, H., Ametz, C., Hetzendorfer, H., Morales, L., Michel, S., & Buerstmayr, H. (2023). Predictor bias in genomic and phenomic selection. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *136*, Article 235. https://doi.org/10.1007/s00122-023-04479-8

Davis, A. P., Mieulet, D., Moat, J., Sarmu, D., & Haggar, J. (2021). Arabica-like flavour in a heat-tolerant wild coffee species. *Nature Plants*, *7*, 413–418. https://doi.org/10.1038/s41477-021-00891-4

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., & Cotes, José M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*, 375–385. https://doi.org/10.1534/genetics.109.101501

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., & Aury, J. M. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, *345*(6201), 1181–1184.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F. (2009). R Software package, e1071. http://cran.rproject.org/web/packages/e1071/index.html

Elias, A. A., Rabbi, I., Kulakow, P., & Jannink, J.-L. (2018). Improving genomic prediction in cassava field experiments using spatial analysis. *G3 Genes, Genomes, Genetics*, *8*, 53–62. https://doi.org/10.1534/g3.117.300323

Fanelli Carvalho, H., Galli, G., Ventorim Ferrão, L. F., Vieira Almeida Nonato, J., Padilha, L., Perez Maluf, M., Ribeiro De Resende, Jr., M. F., Guerreiro Filho, O., & Fritsche-Neto, R. (2020). The effect of bienniality on genomic prediction of yield in arabica coffee. *Euphytica*, *216*, Article 101. https://doi.org/10.1007/s10681-020-02641-7

Ferrão, L. F. V., Dhakal, R., Dias, R., Tieman, D., Whitaker, V., Gore, M. A., Messina, C., & Resende, M. F. R. (2023). Machine learning applications to improve flavor and nutritional content of horticultural crops through breeding and genetics. *Current Opinion in Biotechnology*, *83*, 102968. https://doi.org/10.1016/j.copbio.2023.102968

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Francisco, A., & Garcia, A. A. F. (2017). A mixed model to multiple harvest-location trials applied to genomic prediction in *Coffea canephora*. *Tree Genetics & Genomes*, *13*, Article 95. https://doi.org/10.1007/s11295-017-1171-7

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., & Garcia, A. A. F. (2019). Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models. *Heredity*, *122*, 261–275. https://doi.org/10.1038/s41437-018-0105-y

Ferrão, M. A. G., da Fonseca, A. F. A., Volpi, P. S., de Souza, L. C., Comério, M., Filho, A. C. V., Riva-Souza, E. M., Munoz, P. R., Ferrão, R. G., & Ferrão, L. F. V. (2023). Genomic-assisted breeding for climate-smart coffee. *The Plant Genome*, *17*, e20321.

Ferrão, M. A. G., de Mendonça, R. F., Fonseca, A. F. A., Ferrão, R. G., Senra, J. F. B., Volpi, P. S., Filho, A. C. V., & Comério, M. (2021).

Characterization and genetic diversity of *Coffea canephora* accessions in a germplasm bank in Espírito Santo, Brazil. *Crop Breeding and Applied Biotechnology*, *21*(2), e36132123.

Ferrão, R. G., de Muner, L. H., da Fonseca, A. F. A., & Ferrão, M. A. G. (2019). *Conilon coffee*. Incaper.

Galán, R. J., Bernal-Vasquez, A.-M., Jebsen, C., Piepho, H.-P., Thorwarth, P., Steffan, P., Gordillo, A., & Miedaner, T. (2020). Integration of genotypic, hyperspectral, and phenotypic data to improve biomass yield prediction in hybrid rye. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *133*, 3001–3015. https://doi.org/10.1007/s00122-020-03651-8

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*. https://arxiv.org/pdf/1207.3907

Giraudo, A., Grassi, S., Savorani, F., Gavoci, G., Casiraghi, E., & Geobaldo, F. (2019). Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis. *Food Control*, *99*, 137–145. https://doi.org/10.1016/j.foodcont.2018.12.033

Hernandez, C. O., Wyatt, L. E., & Mazourek, M. R. (2020). Genomic prediction and selection for fruit traits in winter squash. *G3 Genes, Genomes, Genetics*, *10*, 3601–3610. https://doi.org/10.1534/g3.120.401215

Hershberger, J., Morales, N., Simoes, C. C., Ellerbrock, B., Bauchet, G., Mueller, L. A., & Gore, M. A. (2021). Making waves in Breedbase: An integrated spectral data storage and analysis pipeline for plant breeding programs. *The Plant Phenome Journal*, *4*, e20012. https://doi.org/10.1002/ppj2.20012

Hickey, J. M., Chiurugwi, T., Mackay, I., & Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, *49*, 1297–1303. https://doi.org/10.1038/ng.3920

Imbach, P., Fung, E., Hannah, L., Navarro-Racines, C. E., Roubik, D. W., Ricketts, T. H., Harvey, C. A., Donatti, C. I., Läderach, P., Locatelli, B., & Roehrdanz, P. R. (2017). Coupling of pollination services and coffee suitability under climate change. *PNAS*, *114*, 10438–10442. https://doi.org/10.1073/pnas.1617940114

Isik, F., Holland, J., & Maltecca, C. (2017). *Genetic data analysis for plant and animal breeding*. Springer.

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

Krause, M. R., González-Pérez, L., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O., Singh, R. P., Dreisigacker, S., Poland, J., Rutkoski, J., Sorrells, M., Gore, M. A., & Mondal, S. (2019). Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat. *G3 Genes, Genomes, Genetics*, *9*, 1231–1247. https://doi.org/10.1534/g3.118.200856

Lane, H. M., Murray, S. C., Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Rooney, D. K., Barrero-Farfan, I. D., De La Fuente, G. N., & Morgan, C. L. S. (2020). Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. *The Plant Phenome Journal*, *3*, e20002. https://doi.org/10.1002/ppj2.20002

Leroy, T., Montagnon, C., Charrier, A., & Eskes, A. B. (1993). Reciprocal recurrent selection applied to *Coffea canephora* Pierre. I. Characterization and evaluation of breeding populations and value of intergroup hybrids. *Euphytica*, *67*, 113–125. https://doi.org/10.1007/BF00022734

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Mbebi, A. J., Breitler, J.-C., Bordeaux, M., Sulpice, R., Mchale, M., Tong, H., Toniutti, L., Castillo, J. A., Bertrand, B., & Nikoloski, Z. (2022). A comparative analysis of genomic and phenomic predictions of growth-related traits in 3-way coffee hybrids. *G3 Genes, Genomes, Genetics*, *12*, jkac170. https://doi.org/10.1093/g3journal/jkac170

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (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

Mevik, B.-H., & Wehrens, R. (2015). Introduction to the pls package. In *Pls package of R studio software* (pp. 1–23). R Core Team.

Montagnon, C., Leroy, T., Cilas, C., & Charrier, A. (2003). Heritability of *Coffea canephora* yield estimated from several mating designs. *Euphytica*, *133*, 209–218. https://doi.org/10.1023/A:1025543805652

Montagnon, C., Leroy, T., Cilas, C., Legnaté, H., & Charrier, A. (2008). Heterozygous genotypes are efficient testers for assessing between-population combining ability in the reciprocal recurrent selection of *Coffea canephora*. *Euphytica*, *160*, 101–110. https://doi.org/10.1007/s10681-007-9561-9

Mutz, Y. S., Do Rosario, D., Galvan, D., Schwan, R. F., Bernardes, P. C., & Conte-Junior, C. A. (2023). Feasibility of NIR spectroscopy coupled with chemometrics for classification of Brazilian specialty coffee. *Food Control*, *149*, 109696. https://doi.org/10.1016/j.foodcont.2023.109696

Pérez, P., & De Los 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

Posada, H., Ferrand, M., Davrieux, F., Lashermes, P., & Bertrand, B. (2009). Stability across environments of the coffee variety near infrared spectral signature. *Heredity*, *102*, 113–119. https://doi.org/10.1038/hdy.2008.88

Resende, M. F. R. (2016). *High-throughput targeted genotyping of Coffea Arabica and Coffea Canephora using next generation sequencing*. https://api.semanticscholar.org/CorpusID:88274118

Rincent, R., Charpentier, J.-P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., & Segura, V. (2018). Phenomic selection is a low-cost and high-throughput method based on indirect predictions: Proof of concept on wheat and poplar. *G3 Genes, Genomes, Genetics*, *8*, 3961–3972. https://doi.org/10.1534/g3.118.200760

Sousa, T. V., Caixeta, E. T., Alkimim, E. R., Oliveira, A. C. B., Pereira, A. A., Sakiyama, N. S., Zambolim, L., & Resende, M. D. V. (2019). Early selection enabled by the implementation of genomic selection in *Coffea arabica* breeding. *Frontiers in Plant Science*, *9*, 1934. https://doi.org/10.3389/fpls.2018.01934

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

WCR. (2023). *The future of robusta quality: Executive summary*. World Coffee Research. https://worldcoffeeresearch.org/resources/the-future-of-robusta-quality

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, *9*, e1003264. https://doi.org/10.1371/journal.pgen.1003264

Zhu, X., Leiser, W. L., Hahn, V., & Würschum, T. (2021). Phenomic selection is competitive with genomic selection for breeding of complex traits. *The Plant Phenome Journal*, *4*, e20027. https://doi.org/10.1002/ppj2.20027

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Adunola, P., Tavares Flores, E., Riva-Souza, E. M., Ferrão, M. A. G., Senra, J. F. B., Comério, M., Espindula, M. C., Verdin Filho, A. C., Volpi, P. S., Fonseca, A. F. A., Ferrão, R. G., Munoz, P. R., & Ferrão, L. F. V. (2024). A comparison of genomic and phenomic selection methods for yield prediction in *Coffea canephora*. *The Plant Phenome Journal*, *7*, e20109. https://doi.org/10.1002/ppj2.20109