Factor analysis applied in genomic selection studies in the breeding of Coffea canephora

Pedro Thiago Medeiros Paixão · Ana Carolina Campana Nascimento · Moysés Nascimento · Camila Ferreira Azevedo · Gabriela França Oliveira · Felipe Lopes da Silva · Eveline Teixeira Caixeta

Received: 22 April 2021 / Accepted: 23 February 2022 / Published online: 14 March 2022 © The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract Brazil stands out worldwide in the production of coffee. The observed increases in its productivity and morpho agronomic traits are the results of the improvement of several methodologies applied in obtaining improved cultivars, among which the predictive methods of genetic value stand out. These contribute significantly to the selection of higher genotypes, increasing the genetic gain per unit time. In this context, genomic-wide selection (GWS) is a tool that stands out, since it allows predicting the future phenotype of an individual based only on molecular information. Performing joint selection of traits is the interest of most breeding programs, and factor analysis (FA) has been used to assist in this end. The aim of this study was to evaluate the use of FA in the context of GWS, in genotypes of Coffea canephora. It was found that FA was efficient to elucidate the relationships between the traits and generate new variables. The factors formed can assist in the selection, as in addition to allowing joint interpretations, they present good estimates of predictive capacity, heritability and accuracy. Furthermore, high agreement was observed between the individuals selected based on the factors and those selected considering the individual traits. Additionally, it was observed agreement between the top 10% individuals selected based on the “vigor factor” and each variable individually. However, the selection based on “vigor factor” presented individuals with more suitable size from the phytotechnical point of view.

Keywords Genomic prediction · Multivariate analysis · Genetic improvement · Coffea canephora

Introduction

Coffee is one of the most consumed drinks in the world, presenting importance both in terms of its potential health and socioeconomic effects (Porto et al. 2019). The value of this crop has a wide scope, including producer, industry and consumer and consists of the main revenue of many developing countries (Davis et al. 2020; Huded et al. 2020). On the world stage, Brazil stands out as the largest producer of coffee, accounting for about 70% of world exports. Even with the combined effect of the bienniality and the coronavirus crisis, by August/2020 Brazil had already accumulated 3.257 million bags exported in the year (ICO 2020).
Due to the importance of coffee cultivation, breeding programs aim to obtain more productive cultivars, adapted and with better drink quality (Barbosa et al. 2019; Sousa et al. 2017). Thus, in the genetic improvement of the coffee plant, different traits have been worked, such as those associated to plant architecture, adaptation and stability in different cultivation environments and to the size and quality of fruits and grains, as well as resistance and tolerance to biotic and abiotic stress (Marie et al. 2020; Oliveira et al. 2021; Setotaw et al. 2020). In this context, genomic selection (GS) (Meuwissen et al. 2001), has been successfully used in genetic improvement to increase genetic gain per generation through early selection and to improve prediction accuracy (Nascimento et al. 2019). Therefore, GS has an essential role in perennials (Resende et al. 2012; Sousa et al. 2019). The literature presents some applications of GS in coffee. For Coffea arabica, the most cultivated species of the genus, the accuracy of the GS in breeding considering 18 agronomic traits (Sousa et al. 2019) and the effect of biennality on the accuracy of genomic prediction for productivity (Carvalho et al. 2020) were evaluated. For the other cultivated species, C. canephora, the efficiency of GS was analyzed for eight traits of agronomic interest (Alkimim et al. 2020) and the performance of different statistical methodologies were tested within and between environments to predict coffee bean production, rust incidence and green bean yield (Ferrão et al. 2017; 2019).

However, in these studies, GS was applied to individual traits, that is, the results obtained, although promising, are valid only for a single trait. Nevertheless, since these traits are generally correlated, an approach that simultaneously incorporates information from multiple traits into the selection process may be of interest.

One possible method for this approach is factor analysis (FA). FA aims to synthesize the information contained in a set of \( n \) observed variables seeking a new set of \( p \) variables (\( p < n \)), called common latent factors (Macciotta et al. 2012). Thus, the latent variables (common factors) can simultaneously represent a set of traits, without the loss of biological significance, so that the scores of the new variables formed can be treated as pseudo-phenotypes and used in later analyses. In an GWS context, such methodology allows the selection of individuals considering a set of traits simultaneously.

Some studies demonstrate potential applications of FA in genetic improvement. In animals, FA was successfully used to obtain estimates of genetic parameters between common latent factors in cattle and dairy buffaloes (Aspilcueta-Borquis et al. 2012; Macciotta et al. 2006; 2012). Paiva et al. (2019), used formed latent variables, identified as pseudo-phenotypes, in the genetic evaluation of broiler chickens under Bayesian structure. Latent variables also allowed to simultaneously study a set of important characters in pigs, for later use in GS Analyses (Teixeira et al. 2016). For plant species, these studies are still incipient. In the coffee plant, FA was applied to select genotypes of C. arabica with high simultaneous potential of variables of commercial interest (Barbosa et al. 2019). However, despite the importance and considering the best of our knowledge, there are no reports in the literature of the use of FA for genomic prediction in the genus Coffea.

In this context, we aim to: (1) investigate the relationship between several economically important traits in C. canephora through factor analysis; (2) use the latent variables identified as pseudo-phenotypes in genomic prediction; (3) to compare the individuals selected using FA and the individual trait approach.

**Materials and methods**

Phenotypic and genotypic data

Populations composed of clones of C. canephora from the varietal groups Conilon and Robusta and hybrids from crosses between these groups were used. These genotypes make up the breeding program of the Empresa de Pesquisa Agropecuária de Minas Gerais (Epamig) in partnership with the Universidade Federal de Viçosa (UFV) and the Empresa Brasileira de Pesquisa Agropecuária–Café (Embrapa Café). The Conilon genotypes were assigned by the Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (INCAPER) and the Robusta material by the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE) and introduced in the breeding program.

The varietal groups Conilon and Robusta analyzed consisted of 51 and 32 genotypes, respectively. In addition to these genotypes, 82 interpopulational hybrids were obtained through artificial crosses between five
genotypes of the Conilon group and five of the Robusta group.

The genotypes were planted during three consecutive years (2014–2016) following an incomplete block design with up to 35 replicates and single tree plots.

The following morpho agronomic traits were evaluated for three years in a row (2014 and 2016): Vegetative vigor (Vig), evaluated by the overall aspect of the plant, using a scale from 1 to 10, 1 being assigned to the fully-depleted plants and 10 to the highly vigorous plants; Plant height (Phe), measured in centimeters (cm) from ground level to the last apical point of the coffee plant with aid of a measuring tape; Canopy projection diameter (Cdi), given in centimeters (cm) by means of a ruler and in the direction perpendicular to the planting line; and the Production per plant (Prod), evaluated by harvesting all the fruits present in a genotype, and measured the total volume in litres of coffee, freshly harvested.

DNA samples were extracted from young and fully expanded leaves of the same 165 phenotyped coffee plants. The DNA concentration was standardized and sent to RAPiD Genomics, located in Florida, USA, for identification of molecular markers SNPs (Single nucleotide polymorphisms) (Resende et al. 2016). The set of 18,111 identified SNPs markers were submitted to the quality analysis implemented in GemonicLand software (Azevedo et al. 2019). Quality control was achieved using the Kaiser–Meyer–Olkin criterion (KMO) (Ferreira 2011), that is, considering a percentage of explanation of at least 70% of the proportion of the total variance of the traits. To maximize the variability of factor loads and facilitate the interpretation of the distribution of variables in the respective factors, Varimax rotation was used. In this way, the allocation of variables in each factor was made by means of factorial loads, which consist of the correlation between each variable and the respective factors. Thus, the variables that were part of the factor are those most correlated to this.

Subsequently, the factorial scores used were obtained by the regression method, given by:

$$
\hat{Y}_j = \Gamma \hat{F}_j + \hat{c}_j
$$

in which $\Gamma = [y_j]$ is a matrix $(p \times m)$ of coefficients known by rank factor loads $m \leq p$, $F$ is a random vector $(m \times l)$ of non-observable latent common factors and $e$ is a vector $(p \times l)$ of random errors called specific factors (Ferreira 2018).

The adequacy of the correlation structure was evaluated using the Kaiser–Meyer–Olkin criterion (KMO) method and Bartlett test (Cerny and Kaiser 1977). The number of factors $m$ was determined according to (Ferreira 2011), that is, considering a percentage of explanation of at least 70% of the proportion of the total variance of the traits. To maximize the variability of factor loads and facilitate the interpretation of the distribution of variables in the respective factors, Varimax rotation was used. In this way, the allocation of variables in each factor was made by means of factorial loads, which consist of the correlation between each variable and the respective factors. Thus, the variables that were part of the factor are those most correlated to this.

Phenotypic data analysis

The phenotypes were corrected for environmental effects of years and blocks using the software Selegen REML/BLUP (Resende 2016). The model used is equivalent to:

$$
y = Xu + Ta + Wc + Zs + Qb + e,
$$

where $y$ is the vector of phenotypic data; $u$ is the vector of the effects of means of years added to the general mean (assumed to be fixed); $a$ is the vector of additive genetic effects of individuals (assumed to be random and distributed as $a \sim N(0, I \sigma^2_a)$; $c$ is the vector of effects of the specific combining ability between the parents Conilon and Robusta (assumed to be random and distributed as $c \sim N(0, I \sigma^2_c)$; $s$ is the vector of permanent effects of individuals (assumed to be random and distributed as $s \sim N(0, I \sigma^2_s)$; $b$ is the vector of the effects of the permanent environment of the blocks (assumed to be random and distributed as $b \sim N(0, I \sigma^2_b)$, and $e$ is the vector of errors (assumed to be random and distributed as $e \sim N(0, I \sigma^2_e)$. All effects were considered uncorrelated. The incidence matrices of the effects are represented by upper case letters. The phenotypes were adjusted, as de los Campos et al. (2013), considering the following expression $y^* = y - Xu - Qb$.

In order to study several traits simultaneously, the correlation of the corrected phenotypic data set was estimated, and factor analysis was applied in the resulting matrix. The factor analysis model, considering the p-dimensional random vector $Y = [Y_1, Y_2, ..., Y_p]^T$ with mean $\mu$ $(p \times 1)$ and matrix of variances and covariances $\Sigma(p \times p)$ is given by:

$$
Y - \mu = \Gamma F + e
$$

The factor analysis model, considering the p-dimensional random vector $Y = [Y_1, Y_2, ..., Y_p]^T$ with mean $\mu$ $(p \times 1)$ and matrix of variances and covariances $\Sigma(p \times p)$ is given by:

$$
Y - \mu = \Gamma F + e
$$

where $\Gamma = [\gamma_j]$ is a matrix $(p \times m)$ of coefficients known by rank factor loads $m \leq p$, $F$ is a random vector $(m \times l)$ of non-observable latent common factors and $e$ is a vector $(p \times l)$ of random errors called specific factors (Ferreira 2018).

The adequacy of the correlation structure was evaluated using the Kaiser–Meyer–Olkin criterion (KMO) method and Bartlett test (Cerny and Kaiser 1977). The number of factors $m$ was determined according to (Ferreira 2011), that is, considering a percentage of explanation of at least 70% of the proportion of the total variance of the traits. To maximize the variability of factor loads and facilitate the interpretation of the distribution of variables in the respective factors, Varimax rotation was used. In this way, the allocation of variables in each factor was made by means of factorial loads, which consist of the correlation between each variable and the respective factors. Thus, the variables that were part of the factor are those most correlated to this.
in which $\tilde{\Gamma}_{pxm}$ is the matrix of factor loads, $\hat{\psi}_{pxm}$ the matrix of specific factors, $Y_j$ the vector referring to the j-th sample unit and $\hat{\mathbf{Y}}$ the vector of means referring to the phenotypic variables evaluated (Ferreira 2018).

Genomic prediction model

The factor scores calculated for each genotype were used as pseudo-phenotypes (traits) along with the 14,387 SNPs markers in a GWS analysis, in order to estimate the genomic genetic values (GEBVs) of the 165 individuals. To this end, the mixed linear model G-BLUP was used to estimate the individual additive genetic values, according to the expression:

$$y = Xb + Zu_a + \epsilon$$

in which $y$ is the vector of pseudo-phenotypes ($N \times 1$, in which $N$ is the number of individuals); $b$ is the vector of fixed effects ($p \times 1$, in which $p$ is the number of fixed effects), $u_a$ is the vector of additive genetic effects of individuals ($N \times 1$, random), the structures of variances given by $u_a \sim N(0, G_a \sigma_a^2)$ in which $\sigma_a^2$ is the additive variance of the character and $G_a(N \times N)$ is the matrix of genomic kinship between individuals for the additive effects given by $G_a = \frac{\sum_{i,j} q_i q_j}{\sum_i q_i^2}$, where $p_i$ and $q_i$ are the allele frequencies of locus $i$, $W$ is an incidence matrix for the allelic substitution vectors of markers; $e$ is the vector of random residual effects ($N \times 1$) with $e \sim N(0, 1 \sigma_r^2)$ in which $\sigma_r^2$ the residual variance; $X$ is the incidence matrix for fixed effects ($N \times p$); $Z$ is the incidence matrix for random effects ($N \times N$) (Resende et al. 2014). In addition, for comparison purposes, analyses for each measured trait were performed.

After estimating the genetic merit for each interpretable factor and for each trait individually, the genomic genetic correlation between the traits was estimated. Predictive models were also compared using predictive capacity, heritability and accuracy.

Predictive capacity (PC) was estimated from the correlation between the phenotypic values observed for each individual ($\hat{y}$) and their respective GEBVs ($\hat{\mathbf{y}}$), that is: $r_{\hat{y},\hat{\mathbf{y}}} = \frac{\text{Cov}(\hat{\mathbf{y}}, \hat{y})}{\sqrt{\text{Var}(\hat{\mathbf{y}}) \cdot \text{Var}(\hat{y})}}$

In order for the PC not to be overestimated, the cross-validation technique was used. This method allows to evaluate the generalization capacity of a predictive model in a data set (Sousa et al. 2019). The cross-validation method used was $K$-Folds, being considered in this work $K$ equal to 5 folds.

Heritability ($h^2$), which measures the proportion of phenotypic variation in the population attributed to the genetic cause, is given by means of the ratio between genotypic variance $\sigma_g^2$ and phenotypic variance $\sigma_{fn}^2$: $h^2 = \frac{\sigma_g^2}{\sigma_{fn}^2}$.

The accuracy depends on the $h^2$ obtained and the PC, being a measure that is associated with the accuracy in the selection, and is obtained by the estimator (Resende 2015) $r_{q,\hat{y}} = \frac{r_{q,y}}{\sqrt{h^2}}$ in which $r_{q,y} = PC$ and $h^2$ is the heritability.

Based on genetic merit, the best 10% individuals were selected for each approach used. In possession of these individuals, the agreement between the selected individuals was calculated, based on each trait individually and by the factors, using the Cohen’s Kappa coefficient (Cohen 1960). This index can be measured by: $\hat{k} = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)}$,

in which $\text{Pr}(a) - \text{Pr}(e)$ w represents the proportion of observations in which agreement occurred beyond what as expected randomly and $1 - \text{Pr}(e)$, the proportion of observations that no agreement occurred. This measure varies from zero to 1, and the higher the index, the greater the agreement between the groups formed.

Factor analysis was performed in R software (R Core Team, 2019), using the principal function through the psych package (Revelle 2019). Genomic statistical analyses were performed using Genomic-Land software (Azevedo et al. 2019).

**Results**

Factor analysis

The estimated values of the correlation coefficient between the evaluated traits ranged from 0.28 to 0.84 (Fig. 1). According to the $T$ test, at 1% probability, all correlation values were considered significantly different from zero. The value obtained by the KMO statistic was 0.58. In addition, the Bartlett sphericity test was statistically significant ($p-value < 0.01$), indicating that the correlation matrix does not have a diagonal structure, demonstrating the adequacy of the factor analysis.
According to the criterion that the total variation should exceed 70% (Teixeira et al. 2015), two factors with biological interpretations were formed (Tables 1 and 2).

The first factor (Factor 1), that explained about 65% of the original variance (Table 1), grouped traits associated with the morphology of the individual (Vig, Phe and Cdi) (Table 2). This factor has been referred to here as the “vigor factor”. All these traits are positively and highly correlated with the factor (Table 2). Therefore, higher values for these traits will be associated with higher values of the “vigor factor”.

The second factor (Factor 2), which explained about 22% of the variation (Table 1), was constituted only by the production trait (Table 2), this being named as “production factor”.

It was found that the variables belonging to the factors presented values for communalities greater than 0.75, indicating that more than 75% of the variability of these traits was explained by the two factors formed (Table 2).

Accuracy of prediction of traits

The heritability estimates for the evaluated traits were from low to high (Resende, 1997), with values ranging from 0.0034 to 0.6029, respectively (Table 3).

The estimated predictive capacity for the four traits (Vig, Phe, Cdi and Prod) and for the two factors (vigor and production factors) ranged from −0.03 to 0.57 and are presented in Table 3. For these traits, the highest accuracy values were 0.74 and 0.68, obtained for Cdi and Factor 1 respectively (Table 3). Specifically, the value of the PC referring to the “vigor factor” (0.51) is higher when compared to the variables Phe (0.37), Vig (0.39), and is close to the PC of the Cdi trait (0.57) (Table 3). In addition, the accuracy value referring to the “vigor factor” (0.68) is higher when compared to the traits Phe (0.55) and Vig (0.57), which were analyzed individually by the G-BLUP method, and lower when compared to the

---

**Table 1** Eigenvalues of the correlation matrix ($\hat{\lambda}_i$) and percentage of variance explained

| Order (i) | $\hat{\lambda}_i$ | Explained variance (%) | Cumulative variance (%) |
|-----------|------------------|------------------------|------------------------|
| 1         | 2.61             | 65.24                  | 65.24                  |
| 2         | 0.88             | 22.01                  | 87.25                  |
| 3         | 0.43             | 10.80                  | 98.05                  |
| 4         | 0.08             | 1.944                  | 100.0                  |

**Table 2** Observed factors (Factor i), communalities ($\hat{C}$) and uniqueness ($\psi$)

| Variables | Factor 1 | Factor 2 | $\hat{C}$ | $\psi$ |
|-----------|----------|----------|-----------|--------|
| Vig       | 0.89     | 0.04     | 0.80      | 0.20   |
| Phe       | 0.83     | 0.23     | 0.75      | 0.25   |
| Cdi       | 0.97     | 0.14     | 0.95      | 0.05   |
| Prod      | 0.15     | 0.98     | 0.99      | 0.01   |

Vig: vegetative vigor; Phe: plant height; Cdi: canopy projection diameter; Prod: production per plant

---

**Table 3** Heritability $h^2$, predictive capacity (PC) and accuracy for the characteristics and factors formed

| Characteristics | $h^2$  | PC     | Accuracy |
|-----------------|--------|--------|----------|
| Vig             | 0.4899 | 0.3964 | 0.5663   |
| Phe             | 0.4522 | 0.3722 | 0.5535   |
| Cdi             | 0.6029 | 0.5732 | 0.7382   |
| Prod            | 0.0034 | 0.0052 | 0.0892   |
| Factor 1        | 0.5568 | 0.5054 | 0.6773   |
| Factor 2        | 0.0087 | −0.0298| −0.3195  |

Vig: vegetative vigor, Phe: plant height, Cdi: canopy projection diameter, Prod: production per plant, Factor 1: “vigor factor”, Factor 2: “production factor”
variable Cdi (0.74) (Table 3). On the other hand, the estimated predictive capacity for production and Factor 2 (production factor) are similar (Table 3).

Estimates of correlations of genomic genetic values between the four traits (Vig, Phe, Cdi and Prod) and the two formed factors (Factor 1 and Factor 2) ranged from –0.15 to 0.98 (Table 4).

Coefficients of agreement between the top 10% individuals selected based on each variable individually and based on the factors are presented in Table 4. Estimated Cohen’s Kappa values ranged from 0.34 to 0.87. Specifically, considering the first factor (Vigor), only the Apl trait had an estimated Cohen’s Kappa value considered low (0.34). All other estimated Cohen’s Kappa values were greater than 0.50, indicating a very good (Factor 1 x Vig = 0.74; F1 x Cdi = 0.74) and excellent (Factor 2 x Prod = 0.87) rating (Landis and Koch 1977).

The obtained data were also used to calculate the mean, median, standard deviation (SD) and coefficient of variation (CV) for the best 10% genotypes. The best genotypes for each trait were selected individually and considering the “vigor factor”. In general, a reduction in phenotypic values was observed when the selection was performed on the basis of the factor (Table 5).

**Discussion**

In this study, the individual genetic merits of the accesses for four traits (Vig, Phe, Cdi and Prod) and for the two formed factors (vigor and production factors) were predicted by means of the GBLUP model. The joint phenotyping and large-scale genotyping data of C. canephora genotypes were used to evaluate predictive capacity and accuracy through cross-validation with 5 folds (CV). In addition, Cohen’s Kappa correlation and concordance coefficients (based on the highest 10% GEBVs) were estimated among the GEBV values.

The factor analysis was able to satisfactorily represent a set of highly correlated variables, since the KMO value was greater than 0.5 (Hair et al. 2006) and the Bartlett sphericity test was statistically significant ($p$-value $< 0.01$). Since FA is based on a phenotypic correlation matrix, it improves the incorporation of previous knowledge about the underlying genetic

**Table 4** Estimates of the correlation between the genomic genetic values for the evaluated characteristics and the identified factors (upper triangular) and Cohen’s Kappa coefficients (lower triangular)

|        | Vig  | Phe  | Cdi  | Prod | Factor 1 | Factor 2 |
|--------|------|------|------|------|----------|----------|
| Vig    | 1.00 | 0.72**| 0.90**| –0.07* | 0.93**   | –0.15*   |
| Phe    | 1.00 | 0.88**| 0.15* | 0.90**| 0.09*    | 0.09*    |
| Cdi    | 1.00 | 0.08* | 0.98**| 0.98**| 0.00*    | 1.00**   |
| Prod   |      | 1.00 | 0.00* | 0.00**| 1.00     |          |
| Factor 1| 0.74**| 0.34R | 0.74**| 1.00  |          | 0.87E    |
| Factor 2|      |      |       |       | 1.00     |          |

Vig: vegetative vigor; Phe: plant height; Cdi: canopy projection diameter; Prod: production per plant. ** = Significant at 1% probability of error by $T$ test; ns = not significant. E = Excellent, VG = Very Good and R = Reasonable.

**Table 5** Mean, median, standard deviation (SD) and coefficient of variation (CV), for the top 10% genotypes selected for each characteristic individually and based on the “vigor factor”

| Parameter | Phe   | Phe* | Cdi  | Cdi* | Vig  | Vig* |
|-----------|-------|------|------|------|------|------|
| Mean      | 338.8 | 249.2| 264.5| 258.0| 15.20| 13.89|
| Median    | 362.5 | 254.3| 266.0| 266.0| 14.11| 13.93|
| SD        | 93.94 | 95.95| 57.95| 64.65| 4.52 | 5.44 |
| CV        | 27.73%| 38.50%| 21.91%| 25.06%| 29.73%| 39.14%|

Vig: vegetative vigor; Phe: plant height; Cdi: canopy projection diameter; *Vig: vegetative vigor selected based on the "vigor factor"; *Phe: height of the selected plant based on the "vigor factor"; *Cdi: diameter of the selected canopy projection based on the "vigor factor"
architecture, thus allowing the simultaneous selection of characters.

With the exception of the traits production and the factor related to it, the values of heritability were moderate to high (Resende 1997) indicating the possibility of success in the selection based on the evaluated traits (Vig, Cdi, Phe) and the “vigor factor”. The low heritability observed for production indicates that such a trait is controlled by more genes and, therefore, the selection is more complex (Sousa et al. 2019).

The estimation predictive ability and accuracy indicate that it is possible to use the “vigor factor” to predict the genetic merit of genotypes. PC values similar to or higher than those obtained by individual analysis, as obtained in our work, have also been observed in pigs and have been efficiently used in the improvement of this species (Teixeira et al. 2016). On the other hand, the predictive capacity for the trait and the production factor was low. Low values of accuracy and predictive capacity for C. canephora production were reported by Alkimin et al. (2020), corroborating with the results obtained in the present study. In fact, these low values of accuracy and predictive capacity for production are reasonable, since this is a complex trait (Leroy et al. 2011). Although the values were low, the accuracy and predictive capacity of the model can be used to give greater confidence in the evaluation and predicted genetic value of individuals (Resende 2015).

The traits vegetative vigor (Vig), plant height (Phe) and canopy diameter (Cdi) showed positive correlation with the first factor, “vigor factor”. In this case, it is expected that selecting individuals based on the factor means getting taller plants with greater vigor and canopy diameter. This result suggests that this factor may be included as a pseudo-phenotype in C. canephora breeding programs. Ferreira et al. (2005), stressed that the selection of genotypes based on a set of characters of importance is necessary, aiming at simultaneous gains in several traits of agronomic interest.

On the other hand, considering the second factor, “production factor”, since only production presented a high correlation with it, the individual analysis of this trait may be more appropriate. This result is corroborated by those obtained by Ferreira et al. (2005) and Barbosa et al. (2019), which evaluated the possibility of using selection indices in factorial complexes, in characters of genotypes of C. canephora and C. arabica, respectively. These authors also did not group the production trait into any factor.

Positive and significant correlations between GEBVs values (Table 4) were observed between vegetative vigor and canopy projection diameter (0.90), between plant height and canopy diameter (0.88) and between vegetative vigor and plant height (0.72). Since the “vigor factor” represents this set of traits, higher correlation values are observed between them. Given that this estimate captures only the genomic effect, this result also suggests that some genes that influence the expression of these phenotypes may be the same or the occurrence of pleiotropy. In order to detail this result, we can cite the study of genomic association in genotypes of C. canephora of Silva (2018). Using density of 17,885 markers, the presence of SNP with significant associations in the genes for these three traits was observed.

Cohen’s Kappa agreement coefficients (based on the highest 10% GEBVs) were estimated among the GEBV values. According to Landis and Koch (1977), Kappa coefficient values ranging from 0.61 to 0.80 indicate substantial agreement. Therefore, the agreement between the “vigor factor” and the traits Vig (0.74) and Cdi (0.74) can be considered significant. However, the agreement between this factor and Apl (0.34) was considered regular.

In the context of C. canephora breeding, this result may be useful. Specifically, the objective is to identify vigorous genotypes with a larger canopy diameter, traits that are positively correlated to plant height. This behavior constitutes a problem/obstacle in the selection of this crop, since from the phytotechnical point of view, low-sized coffee cultivars are better (Ferreira et al. 2005), as they allow the thickening of plants in the area, providing greater productivity, in addition to reducing the susceptibility of plants to cold winds, facilitating harvesting and cultural treatments (Carvalho 2008). However, as observed by the Kappa coefficient, it is verified that the genotypes selected by means of the "vigor factor" have a lower agreement with those selected considering the plant height trait (Kappa coefficient = 0.34).

In terms of phenotypic values, the selection based on the "vigor factor" allowed to select smaller individuals, with a reduction of 26.46% on average in height. For the other traits (Cdi = 2.45% and Vig = 8.62%), the percentage reduction was lower when compared to that obtained for the Phe trait.
This result is relevant from the phytotechnical point of view, because as discussed earlier, low-sized coffee cultivars are better for allowing the thickening of plants in the area.

The results obtained in the work showed the efficiency of using factor analysis along with genomic selection to predict the individual genetic merits of characters related to the “vigor factor” in the coffee crop. The use of this strategy in the improvement of *C. canephora* is especially important because it is a perennial and long-cycle species and, therefore, the development of new cultivars can take decades (Resende et al. 2012; Sousa et al. 2019). Therefore, the use of genomic selection can speed up the process, making breeding more efficient and competitive. Additionally, the use of the vigor factor can reduce the effect of the correlation between vigor, canopy diameter and plant height, being interesting from the phytotechnical point of view.

**Conclusion**

The factor analysis was efficient for the improvement of *Coffea canephora*, since it allowed to elucidate the structure of relations of the evaluated traits and form new variables, interpreted as "vigor factor" and "production factor". The formed factors presented a satisfactory percentage of explained variability and can be treated as new phenotypes in genomic selection studies. The traits vegetative vigor, plant height and canopy diameter in *C. canephora* can be improved jointly due to its high degree of association. The use of factors proved to be a promising approach associated with GWS for the breeding of *C. canephora*, because in addition to allowing joint interpretations, they were efficient to estimate predictive capacity, heritability and accuracy. In addition, agreement was observed between the top 10% individuals selected based on the “vigor factor” and each variable individually, with the increase in the advantage of the “vigor factor” selecting individuals of more suitable size.

**Acknowledgements** The authors are grateful for the financial support of the Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG (APQ-01638-18), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (Code 001) and the Consórcio Pesquisa Café.

**Declarations**

**Conflict of interest** The authors have no conflict of interest to declare that are relevant to the content of this article.

**Consent for publication** The authors have no financial or proprietary interests in any material discussed in this article.

**References**

Alkimim ER, Caixeta ET, Sousa TV et al (2020) Selective efficiency of genome-wide selection in *Coffea canephora* breeding. Tree Genet Genomes 16:1–11. https://doi.org/10.1007/s11295-020-01433-3

Aspilcueta-Borquis RR, Baldi F, Araujo Neto FR et al (2012) Genetic parameters of total milk yield and factors describing the shape of lactation curve in dairy buffaloes. J Dairy Res 79:60–65. https://doi.org/10.1017/S002209911000823

Azevedo CF, Nascimento M, Fontes VC et al (2019) Genomic prediction. Acta Sci Agron 41:1–7. https://doi.org/10.4025/actasciagreron.v41i1.45361

Barbosa IP, Costa WG, Nascimento M et al (2019) Recommendation of *Coffea arabica* genotypes by factor analysis. Euphytica 215:1–10. https://doi.org/10.1007/s10681-019-2499-x

Carvalho HF, Galli G, Ferrão LFV et al (2020) The effect of bienniality on genomic prediction of yield in arabica coffee. Euphytica 216:1–16. https://doi.org/10.1007/s10681-020-02641-7

Carvalho CHS (2008) (ed.) Cultivares de café. Embrapa, Brasília

Cerný BA, Kaiser HF (1977) A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivar Behav Res 12:43–47. https://doi.org/10.1207/s15327906mbr1201_3

Cohen J (1960) A coeficiente de agrément for nominal scales. Educ Psychol Measur 20:37–46. https://doi.org/10.1177/00131646002000104

Davis AP, Gargiulo R, May MF et al (2020) Lost and found: *Coffea stenophylla* and *C. affinis*, the forgotten coffee crop species of West Africa. Front Plant Sci 11:1–18. https://doi.org/10.3389/fpls.2020.00616

de Los Campos G, Hickey JM, Pont-Wong R, Daetwyler HD, Calus MPL (2013) Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. https://doi.org/10.10344/genetics.112.143313

Ferrão LFV, Ferrão RG, Ferrão MAG et al (2017) A mixed model to multiple harvest-location trials applied to genomic prediction in *Coffea canephora*. Tree Genet Genomes 13:1–13. https://doi.org/10.1007/s11295-017-1171-7

Ferrão LFV, Ferrão RG, Ferrão MAG et al (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
Ferreira A, Cecon PR, Cruz CD et al (2005) Seleção simultânea de Coffea canephora por meio da combinação de análise de fatores e índices de seleção. Pesq Agropec Bras 40:1189–1195. https://doi.org/10.1590/S0100-204X2005001200005

Ferreira DF (2011) Estatística Multivariada. 2.ed. Lavras: Ed. UFLA. 675p.

Ferreira DF (2018) Análise Fatorial Exploratória. In Estatística multivariada. 3.ed.: Ed. UFLA; Lavras, Brasil, pp. 459–507.

Huded AKC, Jingade P, Bychappa M, Mishra MK (2020) Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128. https://doi.org/10.1111/nph.16040

Resende MDV, Silva FF, Azevedo CF (2014) Estatística matemática, biométrica e computacional: Modelos Mistas, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTGWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. Suprema, Viçosa.

Resende M, Caixeta E, Alkimim ER et al (2016) High-throughput targeted genotyping of Coffea arabica and Coffea canephora using next generation sequencing, San Diego, p 1

Resende MDV (1997) Avanços da genética biométrica florestal. In: Encontro Sobre Temas de Genética e Melhoramento, Esaq, Piracicaba, 1997

Resende MDV (2015) Genética quantitativa e de populações. Ltda, Viçosa. 463p

Revelle, W (2019) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://cran.r-project.org/web/packages/psych/index.html

Setotaw TA, Caixeta ET, Zambolim EM et al (2020) Genome introgression of Híbrido de Timor and its potential to develop high cup quality C. arabica cultivars. J Agric Sci 12:64–76. https://doi.org/10.5539/jas.v12n4p64

Silva LF (2018) Estudo de associação genômica ampla (GWAS) em Coffea arabica. Dissertation, Universidade Federal de Viçosa

Sousa TV, Caixeta ET, Alkimim ER et al (2017) Molecular markers useful to discriminate Coffea arabica cultivars with high genetic similarity. Euphytica 213:1–15. https://doi.org/10.1007/s10681-017-1865-9

Sousa TV, Caixeta ET, Alkimim ER et al (2019) Early selection enabled by the implementation of genomic selection in Coffea arábica breeding. Front Plant Sci 9:1–12. https://doi.org/10.3389/fpls.2018.01934

Teixeira FRF, Nascimento M, Nascimento ACC et al (2015) Determinação de fatores em características de suínos. Rev Bras Biom 33:130–138

Teixeira FRF, Nascimento M, Nascimento ACC et al (2016) Factor analysis applied to genome prediction for high-dimensional phenotypes in pigs. GMR 15:1–10. https://doi.org/10.3389/gmr.15028231

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.