Tree Genetics & Genomes (2023) 19:23
https://doi.org/10.1007/s11295-023-01597-8

ORIGINAL ARTICLE

Genome-wide association study for morphological, physiological, and productive traits in *Coffea arabica* using structural equation models

Matheus Massariol Suela[1] · Camila Ferreira Azevedo[1] · Ana Carolina Campana Nascimento[1] · Mehdi Momen[2] · Antônio Carlos Baião de Oliveira[3] · Eveline Teixeira Caixeta[3] · Gota Morota[4] · Moysés Nascimento[1]

Received: 1 November 2022 / Revised: 1 March 2023 / Accepted: 17 March 2023

Abstract

Yield is one of the most important traits of arabica coffee. Plant breeders seek to maximize yield directly or indirectly, using other related traits. The standard multi-trait genome-wide association study (MTM-GWAS) does not accommodate the network structure of phenotypes, therefore, does not address how traits are interrelated. We applied structural equation modeling (SEM) to GWAS to explore interrelated dependencies between phenotypes related to morphology (fruit size and number of reproductive nodes), physiology (vegetative vigor), and productivity (yield) traits using 195 *Coffea arabica* individuals genotyped with 21,211 single-nucleotide polymorphism markers. We inferred the probabilistic phenotypic network by the Hill-Climbing algorithm to estimate the structural coefficients. The integration of multivariate GWAS and SEM (SEM-GWAS) identified a positive interrelationship between vegetative vigor and yield, and vegetative vigor and the number of reproductive nodes. Among those traits, yield and number of reproductive nodes presented indirect SNP effects. There was no evidence of a single quantitative trait locus controlling all the traits jointly. We identified three genes (Stress enhanced protein 1, Abscisic stress-ripening protein 5, and SAR–SNI1) that acted directly on yield. In summary, SEM-GWAS offered new insights into the relationship between the traits linked to coffee yield, providing useful information for arabica coffee breeding programs.

Keywords Structural equation model · Bayesian network · Genome-wide association study · Single-nucleotide polymorphism · *Coffea arabica*

Introduction

Coffee is one of the most widely consumed beverages worldwide, which Brazil being the world's largest producer. Brazil produces 39.76% of all the coffee in the world (*Coffea canephora* and *Coffea arabica*). In particular, Brazil accounts for 40.83% of Arabica coffee worldwide (Estados Unidos 2021). Due to the increase in coffee consumption in countries that are not as traditional, such as China (DCCC 2019), climate change risks, the demand for coffees with ever-increasing beverage quality (Borém et al. 2021), among others, it is necessary to encourage targeted research studies of breeding strategies so that greater sustainability of the production chain can be achieved. Genetic-driven breeding is a tool that an enable such advances to meet market demands (Barka et al. 2017; Wallace et al. 2018). However, the breeding process

Responsible Editor: C. Kulheim.

✉ Matheus Massariol Suela
massariolsuela97@gmail.com

[1] Department of Statistics, Federal Univesity of Viçosa, Viçosa, MG, Brazil

[2] Department of Surgical Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, USA

[3] Embrapa Coffee, Brazilian Agricultural Research Corporation (Embrapa), Brasília, Brazil

[4] School of Animal Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Published online: 12 April 2023

takes time because Arabica coffee has a long cycle and juvenile period (Ferrão et al. 2016; Nonato et al. 2021). Thus, the integration of innovative tools, such as biotechnology coupled with quantitative genetic approaches and genetic-driven breeding, is needed to make the genetic progress of Arabica coffee and enable such advances to meet market demands (Ferrão et al. 2016; Nonato et al. 2021; Mishra and Slater 2012).

Genome-wide association studies (GWAS) have become increasingly popular for elucidating the genetic architecture of economically important traits (Yu et al. 2006). In coffee, GWAS have been successfully used in identifying regions in the genome associated with essential traits, such as rust resistance (Romero et al. 2014; Sousa et al. 2020), fruit size, yield, plant height (Sousa et al. 2020), lipid biochemistry and diterpene content Sant'Anna et al. (2018), caffeine biosynthesis (Tran et al. 2018), and resistance to coffee berry disease (Gimase et al. 2020). Generally, methodologies developed in GWAS consider each trait individually. However, typically correlated traits are recorded in the same material in breeding programs. The univariate approach may be ineffective in examining the genetic interdependence of traits and may impose limitations on elucidating the genetic mechanisms underlying a complex system among traits (Momen et al. 2019). As an alternative, multivariate GWAS models (MTM-GWAS) can reduce the false positive rate and increases the statistical power of association tests (Zhou and Stephens 2012; Korte et al. 2012; O'Reilly et al. 2012). This approach allows the identification of genomic regions with pleiotropic effects explaining genetic correlation among the traits. Although MTM-GWAS is a useful approach, this methodology does not include how the traits are interrelated.

Some methodologies that deal with GWAS in a multivariate way, such as mvBIMBAM methodology (Shim et al. 2015), which is based on the Bayes factor, partitioning the effects of markers directly and indirectly, in addition to this, Momen et al. (2018) proposed the use of structural equation modeling to perform MTM-GWAS (SEM-GWAS) and applied it to crop plants (Momen et al. 2019). This model was later extended to a Bayesian marker effect model (Wang et al. 2020). The SEM-GWAS approach captures complex relationships and delivers a more comprehensive understanding of single-nucleotide polymorphism (SNP) effects than MTM-GWAS. Specifically, it can partition the total SNP effects of a trait into direct and indirect effects, enhancing our understanding of complex relationships among agronomic traits. Furthermore, SEM-GWAS has the potential to provide deeper insights into the underlying genetic architecture of important traits in breeding programs than what is currently possible with MTM-GWAS. Thus, our objectives were to (1) estimate genetic parameters for phenological traits in *Coffea arabica*; and (2) enhance the understanding of the genetic architecture of agronomic traits using the SEM-GWAS approach.

Materials and methods

Phenotypic and genotypic data

The data were collected from the *C. arabica* breeding program, which is a joint partnership among the Company of Farming Research of Minas Gerais (EPAMIG), the Federal University of Viçosa (UFV), and the Brazilian Agricultural Research Corporation (EMBRAPA). An experimental area is maintained at the Department of Phytopathology—UFV (lat. 20°44′25″ S, long. 42°50′52″ W). The database contains 13 progenies from crosses between three parents of the Catuaí cultivar and three parents of the Híbrido de Timor (HdT). Fifteen full-sib families of progeny mentioned above, totaling 195 individuals, were genotyped with m = 21,211 SNP markers. The DNA concentration of the samples was standardized and sent to RAPiD GENOMICS, Florida/USA, for probes construction, sequencing, and identification of SNP molecular markers (Sousa et al. 2017). The SNP quality control was carried out considering genotypic call rate and minor allele frequency equal to or greater than 90% and 5%, respectively.

The genotypes were planted on February 11, 2011, using the spacing of 3.0 m between rows and 0.7 m between plants. Nutritional management was carried out following the requirements of the crop. From the cross between three parents of the Catuaí group and three parents of Híbrido de Timor (HdT), which contrast in relation to resistance to coffee rust, 13 progenies were obtained from the *C. arabica* breeding program of Epamig/UFV/Embrapa. These progenies are resistant backcrosses (BCr), susceptible backcrosses (BCs), and F_2 generations. Thirteen progenies, which were composed of 15 plants (repetitions), were analyzed, totaling 195 individuals. The field trial included experiments in different years, with blocks and plot randomization. More details can be found in Sousa et al. (2019). The phenotypic database used was comprised of four traits: yield (YL), vegetative vigor (VV), number of reproductive nodes (NRN), and fruit size (FS). The importance of these traits has been reported in the literature. For example, according to Cilas et al. 2006, individuals with larger amounts of NRN tend to have higher YL. Ferrão et al. 2012 showed that FS is one of the main traits used to select production performance, and VV indicates the growth potential. Finally, the main target trait in the breeding program is YL, which is influenced by other traits, including VV, NRN, and FS.

Phenotypic modeling

The phenotypic values of the traits were adjusted for years (2014, 2015, and 2016), plots, and years × plots interaction. The analyses were performed based on a mixed linear model (REML/BLUP) using the Selegen-REML/BLUP software [http://www.ppestbio.ufv.br/wp-content/uploads/2016/05/Software_Selegen_Genomica.zip] (Resende 2016). The statistical model was

$$\mathbf{y} = \mathbf{Xu} + \mathbf{Zg} + \mathbf{Wp} + \mathbf{Vr} + \mathbf{Tb} + \mathbf{Ri} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the long vector of phenotype concatenated together, $\mathbf{u}$ is the vector of general average in each year of evaluation (fixed effect), $\mathbf{g}$ is the vector of progeny effects (random effect), $\mathbf{p}$ is the vector of permanent effects between individuals (random effect), $\mathbf{r}$ is the effects between backcross and F_2 population (random effect), $\mathbf{b}$ is the effects between plot (random effect), $\mathbf{i}$ is the progenies x years interaction effect (random effect), and $\mathbf{e}$ is a vector of model residuals (random effect) (Sousa et al. 2019). Here, $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{V}$, $\mathbf{T}$, and $\mathbf{R}$ are the incidence matrices related to $\mathbf{g}$, $\mathbf{p}$, $\mathbf{r}$, $\mathbf{b}$, $\mathbf{i}$, and $\mathbf{e}$, respectively and were assumed to follow a normal distribution $\mathbf{g} \sim N(0, \mathbf{I}\sigma_g^2)$, $\mathbf{p} \sim N(0, \mathbf{I}\sigma_p^2)$, $\mathbf{r} \sim N(0, \mathbf{I}\sigma_r^2)$, $\mathbf{b} \sim N\left(0, \mathbf{I}\sigma_b^2\right)$, and $\mathbf{i} \sim N(0, \mathbf{I}\sigma_i^2)$, where $\mathbf{I}$ is the identity matrix. The adjusted phenotypes are given by $\mathbf{y}^* = \mathbf{Z}\hat{\mathbf{g}} + \hat{\mathbf{e}}$ (de Los Campos et al. 2013).

Bayesian multi-trait genomic best linear unbiased prediction

The adjusted phenotypes ($\mathbf{y}^*$) were used as input in a Bayesian multi-trait genomic best linear unbiased prediction model, which can be described as follows:

$$\mathbf{y}^* = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e} \quad (2)$$

where $\mathbf{y}^*$ is the long vector of adjusted phenotypes concatenated together (YL, VV, FS, and NRN, t = four traits), $\mathbf{X}$ is the incidence matrix of non-genetic effects, in this case the general mean; $\mathbf{b}$ is the vector of the non-genetic effects; $\mathbf{Z}$ is the incidence matrix relating phenotypes and progenies; $\mathbf{g}$ is the vector of additive genetic effect; and $\mathbf{e}$ is the vector of model residuals. The $\mathbf{g}$ and $\mathbf{e}$ vectors were assumed to follow multivariate Gaussian distribution $\mathbf{g} \sim N(0, \Sigma_g \otimes \mathbf{G})$ and $\mathbf{e} \sim N(0, \Sigma_e \otimes \mathbf{I})$, respectively, where $\mathbf{G}$ is the genomic relationship matrix, $\mathbf{I}$ is the identity matrix, $\mathbf{\Sigma}_g$ and $\mathbf{\Sigma}_e$ are the t × t variance–covariance matrices of genetic effect and residuals, respectively, and ⊗ indicates the Kronecker product. The $\mathbf{G}$ matrix was computed as $\mathbf{G} = \mathbf{WW}'/2\sum_{j=1}^{m} p_j(1-p_j)$, where $\mathbf{W}$ is the centered SNP marker matrix (VanRaden 2008). A flat prior was assigned to the vector of $\mathbf{b}$. An inverse Wishart distribution, with hyperparameters ν (degree of freedom) and S (scale parameter), was assumed for the covariances matrices $\mathbf{\Sigma}_g$ and $\mathbf{\Sigma}_e$. The significance of the genetic correlations, residuals, and heritabilities was based on the relative highest 95% probability density (HPD 95%) intervals, where intervals containing zero it was concluded that the estimator is equal to zero. For intervals that do not contain zero, the estimator is non-zero.

Marginal posterior densities were obtained using a Markov Chain Monte Carlo (MCMC) method using the Gibbs sampling algorithm, with 1,200,000 MCMC iterations, a burn-in of 50,000, and a thinning rate of 50 resulting in 23,000 MCMC samples for inference. The posterior means of the residuals were used as inputs for the following Bayesian network analysis to infer a trait network. The convergence analysis is presented on Tables S2 to S9. The autocorrelations (Tables S2 and S6) and convergence tests were done via Geweke (Tables S3 and S7), Heidleberger and Welch Stationarity and Interval Halfwidth (Tables S4 and S8), and Raftery and Lewis (Tables S5 and S9) for the residual and genetic chains of the model, respectively.

Bayesian networks

Bayesian networks are graphical models, where nodes represent random variables (phenotypes), and edges represent probabilistic dependencies between them (Korb and Nicholson 2010). The Hill Climbing (HC) score-based algorithm was used to construct a potential interrelationship among traits as implemented in the bnlearn R package (Scutari 2010). The Bayesian information criterion (BIC) was computed for each edge removal to infer their relative contribution to the overall BIC of the network. We estimated the strength and uncertainty of the direction of each edge probabilistically by bootstrapping (50,000 bootstrap samples). An edge strength ≥ 80% was used to select high-confidence relationships.

The percentages reported adjacent to the edges and in parentheses indicate the proportion of the bootstrap samples supporting the edge (strength) and the direction of the edge, respectively. Strength is the probability that an arc exists between the nodes, independently of its direction. Direction is the probability that an arc has a certain direction. YL: yield; VV: vegetative vigor; FS: fruit size; NRN: number of reproductive nodes.

Multi-trait association analysis (MTM-GWAS)

MTM-GWAS analyses were performed using the SNP Snappy strategy (Meyer and Tier 2012) implemented in the mixed model WOMBAT program (Meyer 2007). This model does not consider the inferred network structure:

$$\mathbf{y}^{**} = \mathbf{W}_j\mathbf{s}_j + \mathbf{Zg} + \mathbf{e} \quad (3)$$

where $\mathbf{y}^{**}$ is the long vector of scaled phenotypes concatenated together (t = 4), $\mathbf{W}_j$ is the matrix of SNP codes for the j th marker, $\mathbf{s}_j$ is the vector of j th SNP marker effect, and other terms were previously described. Variance–covariance structures were assumed the same as for Eq. (2). The MTM-GWAS was fitted for each SNP individually to obtain the marker effects for each trait, i.e., $\mathbf{s} = [s_{YL}, s_{VV}, s_{FS}, s_{NRN}]$. A t statistic was used to obtain $p-$ values: $t_{ij} = s_j/se(s_j)$, where s is the point estimate of the j th SNP effect and $se(s_j)$ is its standard error. The q values were obtained by correcting the p values (Storey and Tibshirani 2003). In addition to the previously mentioned corrections, the SNP dosage matrix was also corrected for the population structure, where 4 principal components were used, resulting in approximately 80% of the genetic variability.

Structural equations modeling GWAS

The structural equation modeling (SEM) incorporates a trait network structure into the GWAS framework. Structural equations modeling GWAS (SEM-GWAS) was conducted using the SNP Snappy strategy (Meyer and Tier 2012). The SEM described in Gianola and Sorensen (2004) was extended to GWAS (Momen et al. 2019, 2018).

$$\mathbf{y}^{**} = \mathbf{\Lambda y}^{*} + \mathbf{Ws} + \mathbf{Zg} + \mathbf{e} \quad (4)$$

where $\mathbf{\Lambda}$ is a matrix of structural coefficients based on the learned structure from the Bayesian network:

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ I\lambda_{VV\to NRN} & 0 & 0 & 0 \\ I\lambda_{VV\to FS} & I\lambda_{NRN\to FS} & 0 & 0 \\ I\lambda_{VV\to YL} & I\lambda_{NRN\to YL} & I\lambda_{FS\to YL} & 0 \end{bmatrix}$$

The vectors $\mathbf{g}$ and $\mathbf{e}$ were assumed to have a joint distribution $\begin{bmatrix} \boldsymbol{g} \\ \boldsymbol{e} \end{bmatrix} = N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_g \otimes \mathbf{G} & 0 \\ 0 & \boldsymbol{\psi} \end{bmatrix} \right\}$, and the residual covariance matrix was diagonal with

$$\psi = \begin{bmatrix} \sigma^2_{e(VV)} & 0 & 0 & 0 \\ 0 & \sigma^2_{e(NRN)} & 0 & 0 \\ 0 & 0 & \sigma^2_{e(FS)} & 0 \\ 0 & 0 & 0 & \sigma^2_{e(YL)} \end{bmatrix}$$

Using the results provided by the Bayesian network, the interrelationships between the four traits (YL, VV, NRN, and FS) were modeled using SEM. The structural equations to estimate SEM parameters and SNP effects can be expressed as follows:

$$\boldsymbol{y}_{1_{YL}} = \lambda_{21}y_{2_{VV}} + \lambda_{31}y_{3_{FS}} + \lambda_{41}y_{4_{NRN}} + \lambda_{41}\lambda_{21}y_{4_{NRN}} + \boldsymbol{w}_j s_{j(y1_{YL})} + \boldsymbol{Z}_1\boldsymbol{g}_1 + \boldsymbol{\epsilon}_1$$

$$= \lambda_{21}[\boldsymbol{w}_j s_{j(y2_{VV})} + \boldsymbol{Z}_2\boldsymbol{g}_2 + \boldsymbol{\epsilon}_2] + \lambda_{31}\left[\boldsymbol{w}_j s_{j(y3_{FS})} + \boldsymbol{Z}_3\boldsymbol{g}_3 + \boldsymbol{\epsilon}_2\right] + \lambda_{41}[\lambda_{24}(\boldsymbol{w}_j s_{j(y2_{VV})} + \boldsymbol{Z}_2\boldsymbol{g}_2 + \boldsymbol{\epsilon}_2) + \boldsymbol{w}_j s_{j(y4_{NRN})} + \boldsymbol{Z}_4\boldsymbol{g}_4 + \boldsymbol{\epsilon}_4] + \boldsymbol{w}_j s_{j(y1_{YL})} + \boldsymbol{Z}_1\boldsymbol{g}_1 + \boldsymbol{\epsilon}_1$$

$$\boldsymbol{y}_{2_{VV}} = \boldsymbol{w}_j s_{j(y2_{VV})} + \boldsymbol{Z}_2\boldsymbol{g}_2 + \boldsymbol{\epsilon}_2$$

$$\boldsymbol{y}_{3_{FS}} = \boldsymbol{w}_j s_{j(y3_{FS})} + \boldsymbol{Z}_3\boldsymbol{g}_3 + \boldsymbol{\epsilon}_3$$

$$\boldsymbol{y}_{4_{NRN}} = \lambda_{24}y_{2_{VV}} + \boldsymbol{w}_j s_{j(y4_{NRN})} + \boldsymbol{Z}_4\boldsymbol{g}_4 + \boldsymbol{\epsilon}_4$$

$$= \lambda_{24}[\boldsymbol{w}_j s_{j(y2_{VV})} + \boldsymbol{Z}_2\boldsymbol{g}_2 + \boldsymbol{\epsilon}_2] + \boldsymbol{w}_j s_{j(y4_{NRN})} + \boldsymbol{Z}_4\boldsymbol{g}_4 + \boldsymbol{\epsilon}_4$$

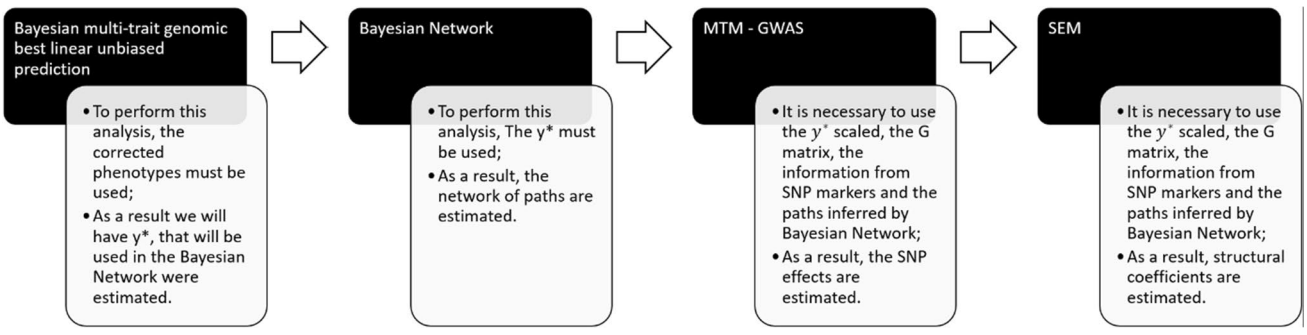

Fig. 1 Flowchart detailing the inputs and outputs of Bayesian multi-trait genomic best linear unbiased prediction, Bayesian network, multi-trait genome-wide association study (MTM-GWAS), and structural equations modeling genome-wide association study (SEM-GWAS)

Thus, the corresponding estimated $\mathbf{\Lambda}$ matrix was

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ I\lambda_{VV\rightarrow NRN} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ I\lambda_{VV\rightarrow YL} & I\lambda_{NRN\rightarrow YL} & I\lambda_{FS\rightarrow YL} & 0 \end{bmatrix}$$

The SEM-GWAS procedures are summarized in Fig. 1.

The effect sizes of edges between phenotypes in the Bayesian network were estimated as the structural coefficients. They were used to develop a set of structural equations to partition the total SNP effects into direct and indirect components. While MTM-GWAS estimates the effect of SNP as a total effect, SEM further partitions it into the direct and indirect effects of SNP, considering the trait network. The calculation of indirect effects is based on the multiplication of path coefficients for each path linking the SNP to an associated variable and then adding all these paths (Momen et al. 2019, 2018). The knowledge of direct and indirect effects is of great importance for understanding total SNP effects, which is not possible using MTM-GWAS (Momen et al. 2019, 2018; Valente et al. 2013).

Pathway enrichment analyses

We first select relevant SNP from the MTM-GWAS results based on a nominal *P* value (< 0.01). The QTLs were associated with positions without considering window. Then, we used the genomic database of *Coffea arabica* available at the NCBI (National Center for Biotechnology Information—https://www.ncbi.nlm.nih.gov/genome/?term=txid13443[Organism:noexp]) to estimate the functionality of each gene identified and the significantly enriched GO terms and KEGG pathways.

Results

Descriptive statistics for the traits investigated are reported in Table 1. Their averages were 5.19 ± 0.21 L/plant (4.77, 5.59) for YL, 7.35 ± 2.63 (1.99, 7.47) for VV, 2.32 ± 0.14 (1.99, 2.37) for FS, and 8.62 ± 0.65 (7.19, 8.89) for NRN.

Table 1 Means, HPD95 (highest 95% probability density) intervals and standard deviations (SD) for yield (YL), vegetative vigor (VV), fruit size (FS), and number of reproductive nodes (NRN) measured in 195 *C. arabica* genotypes

Trait	Mean	SD	HPD_{lower}	HPD_{upper}
YL	5.19	0.21	4.77	5.59
VV	7.35	2.63	1.99	7.47
FS	2.32	0.14	1.99	2.37
NRN	8.62	0.65	7.19	8.89

Structure analysis

After carrying out the principal component analysis, a structure was detected in which four principal components explained 80.07% of the data variability. The first four components explained 42.82%, 30.90%, 3.62%, and 2.73%, respectively. The figure corresponding to the above analysis for two principal components can be seen below (Fig. 2).

Genetic parameters

The estimates of narrow sense heritability were moderate for VV (0.39 ± 0.14) and FS (0.61 ± 0.12) and small for YL (0.14 ± 0.10) and NRN (0.13 ± 0.17) (Table 2). No genomic correlation was considered significant according to the highest 95% probability density intervals (Table 2). We found relevant positive correlations among residual correlations between FS and YL (0.30) and NRN and YL (0.38) (Table 2).

Bayesian network structure

The Bayesian network analysis showed that there is a directed path from VV to NRN (68% of bootstrap samples and 100% of strength), NRN to YL (100% of bootstrap samples and 81% of strength), VV to YL (100% of bootstrap samples and 99% of strength), and FS to YL (100% of bootstrap samples and 82% of strength). The NRN mediated the indirect path between VV and YL (Fig. 3). In the inferred trait network, YL was downstream, VV and FS were upstream, and NRN was the mediator trait.

The BIC was used as a goodness-of-fit statistic, which measures how well the paths mirror the dependence structure of the data. The greatest decrease in BIC was observed when removing the VV → NRN (− 15.19) and VV → YL (− 15.09) paths, suggesting that these paths may play the most important role in the trait network (Table 3). The NRN → YL and FS → YL paths presented BIC equal to − 2.37 and − 2.54, respectively, showing that these are paths that would not have such an impact, if removed, on the fit of the model when compared to the first two.

Structural equation model

The estimated structural coefficients are shown in Table 4. The coefficients of the NRN → YL and FS → YL paths were negative, while VV → NRN and VV → YL were positive. The coefficient referring to the NRN → YL path had the highest value, while VV → YL had the lowest coefficient. However, all the coefficients were small, suggesting that the role of VV in mediating SNP effects on NRN and YL and the role of NRN and FS in mediating SNP effects on YL are marginal. The magnitude of structural coefficients suggested

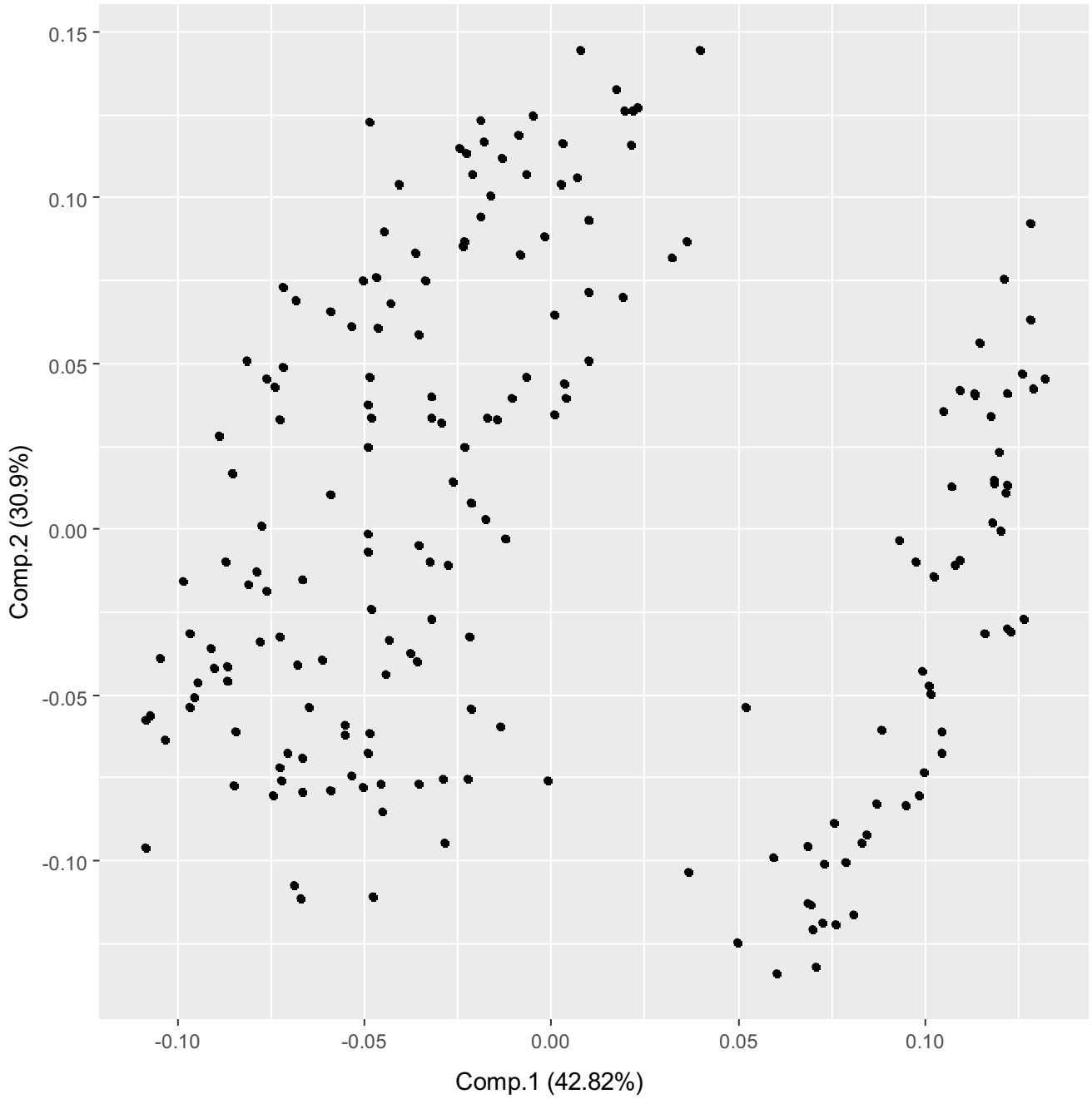

Fig. 2 Principal components analysis of *Coffea arabica* panel. Scatter plot of the first two principal components (Comp.1 and Comp.2)

Table 2 Genomic (upper triangular) and residual (lower triangular) correlations, and genomic heritabilities (diagonal) and their respective HPD95 (highest 95% probability density) in parenthesis for yield (YL), vegetative vigor (VV), fruit size (FS), and number of reproductive nodes (NRN). Significant correlations (HPD95 not including 0) are highlighted in bold

	YL	VV	FS	NRN
YL	0.14 (0.01, 0.33)	0.44 (−0.64, 0.92)	−0.32 (−0.81, 0.51)	0.57 (−0.49, 0.98)
VV	0.47 (−0.23, 0.58)	0.39 (0.13, 0.66)	−0.30 (−0.72, 0.64)	0.40 (−0.67, 0.90)
FS	**0.30 (0.03, 0.45)**	−0.01 (−0.17, 0.28)	0.61 (0.33, 0.79)	−0.25 (−0.79, 0.49)
NRN	**0.38 (0.27, 0.59)**	0.40 (−0.25, 0.53)	0.19 (−0.17, 0.35)	0.13 (0.01, 0.56)

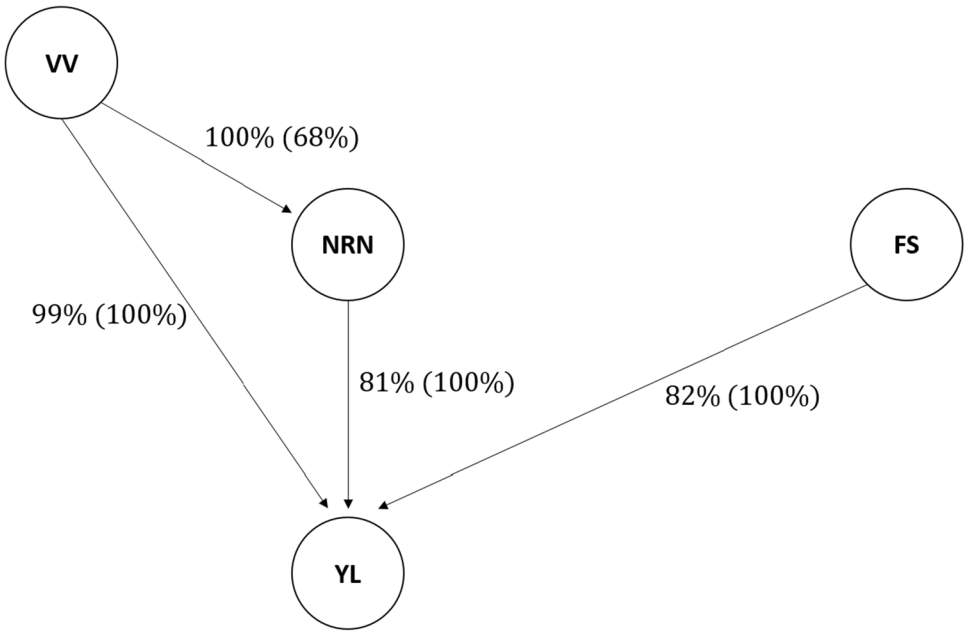

Fig. 3 The trait network structure inferred by the Hill-Climbing algorithm. The structure learning test was performed with 50,000 bootstrap samples. The percentages reported adjacent to the edges and in parentheses indicate the proportion of the bootstrap samples supporting the edge (strength) and the direction of the edge, respectively. Strength is the probability that an arc exists between the nodes, independently of its direction. Direction is the probability that an arc has a certain direction. *YL* yield, *VV* vegetative vigor, *FS* fruit size, *NRN* number of reproductive nodes

Table 3 Bayesian information criterion (BIC) of the network learned using the Hill-Climbing algorithm

BIC (a)	Path	BIC (b)
− 1395.29	VV → NRN	− 15.19
	VV → YL	− 15.09
	NRN → YL	− 2.37
	FS → YL	− 2.54

(a) Bayesian information criterion score (BIC) of the entire network; (b) BIC scores for pairs of nodes; the change in the score when removing the arc relative to the entire network score is shown. *YL* yield, *VV* vegetative vigor, *FS* fruit size, *NRN* number of reproductive nodes

Table 4 Structural coefficients (λ) estimates derived from the structural equation model

Path	Path coefficient (λ)
VV → NRN	0.0351
VV → YL	0.0047
NRN → YL	− 0.0438
FS → YL	− 0.0338

YL yield, *VV* vegetative vigor, *FS* fruit size, *NRN* number of reproductive nodes

that allelic substitutions in quantitative trait loci (QTL) for one trait might affect another trait. The positive coefficient λ_{21} (0.0047) quantifies the effect of the VV in YL and λ_{24} (0.0351) quantifies the effect of the VV in NRN. This suggests that a 1-unit increase in VV results in a 0.0047 unit and 0.0351 unit increase in YL and NRN, respectively. Likewise, the negative effects λ_{31} and λ_{41} offer the same interpretation. The negative coefficients λ_{41} (−0.0438) and λ_{31} (−0.0338) quantify the effects of NRN in YL and FS in YL, respectively, suggesting that a 1-unit increase in NRN results in a 0.0438 unit decrease in YL and a 1-unit increase in FS results in a 0.0338 unit decrease in YL.

Figure 4 shows a graphical representation of the trait network and SNP effects using the structural equation modeling (SEM) model. Here, we can observe the direct effects of markers of YL [$s_{j(YL)}$], VV [$s_{j(VV)}$], NRN [$s_{j(NRN)}$], and FS [$s_{j(FS)}$], and also the indirect effects between VV and NRN (λ_{24}), VV and YL (λ_{21}), NRN and YL (λ_{41}), and FS and YL (λ_{31}).

Partitioning of SNP effects

Using SEM-GWAS, we partitioned the effects of SNP into direct and indirect effects. The results of the decomposition of SNP effects are presented for each trait separately.

Yield

The overall SNP effect of yield (YL) was partitioned into one direct and four indirect effects, according to Fig. 4. Specifically, VV, NRN, and FS affected YL through indirect paths with structural coefficients equal to $\lambda_{21} = 0.0047$, $\lambda_{41} = -0.0438$, and $\lambda_{31} = -0.0338$, respectively. VV also indirectly contributed to NRN, which in turn affected YL, represented by the product of the coefficients λ_{41}x λ_{24} ($-0.0438 \times 0.0351 = -0.0015$). The total effect of *j* th SNP on YL was equal to the sum of the direct and indirect effects.

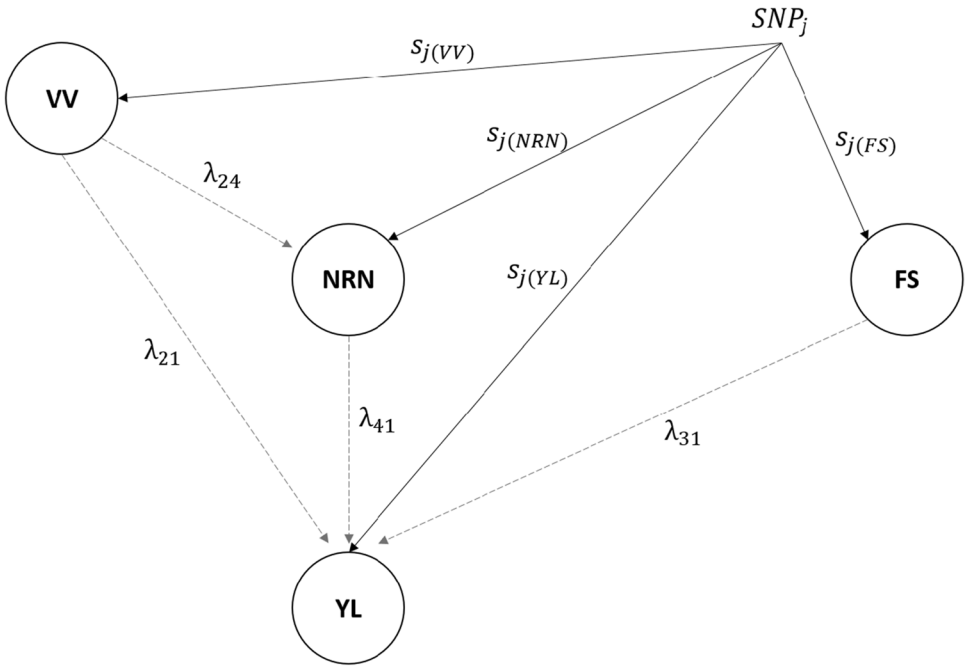

Fig. 4 Path analysis of marker effects based on the inferred trait network. *YL* yield, *NRN* number of reproductive nodes, *FS* fruit size, *VV* vegetative vigor. The gray dashed arrows indicate the direction of relationships. λ_{24}: VV → NRN; λ_{21}: VV → YL; λ_{41}: NRN → YL; λ_{31}: FS → YL. The black arrows indicate the direct effect of the j th SNP

$$\text{Direct}_{s_j \to y1_{YL}} = s_{j(y1_{YL})}$$

$$\text{Indirect}(1)_{s_j \to y1_{YL}} = \lambda_{21} s_{j(y2_{VV})}$$

$$\text{Indirect}(2)_{s_j \to y1_{YL}} = \lambda_{41} s_{j(y4_{NRN})}$$

$$\text{Indirect}(3)_{s_j \to y1_{YL}} = \lambda_{31} s_{j(y3_{FS})}$$

$$\text{Indirect}(4)_{s_j \to y1_{YL}} = \lambda_{41}\lambda_{24} s_{j(y2_{VV})}$$

$$\begin{aligned}\text{Total}_{s_j \to y1_{YL}} &= \text{Direct}_{s_j \to y1_{YL}} \text{Indirect}(1)_{s_j \to y1_{YL}} \\ &+ +\text{Indirect}(2)_{s_j \to y1_{YL}} + \text{Indirect}(3)_{s_j \to y1_{YL}} \\ &+ \text{Indirect}(4)_{s_j \to y1_{YL}} = s_{j(y1_{YL})} + \lambda_{21} s_{j(y2_{VV})} \\ &+ \lambda_{41} s_{j(y4_{NRN})} + \lambda_{31} s_{j(y3_{FS})} + \lambda_{41}\lambda_{24} s_{j(y2_{VV})}\end{aligned}$$

The Manhattan plots of direct (A), indirect (B), and total (C) SNP effects on YL are presented in Fig. 5.

Number of reproductive nodes

The SNP effect on number of reproductive nodes (NRN) was decomposed into one direct and one indirect effect mediated by VV (VV → NRN) with the structural coefficient λ_{24} (0.0351).

$$\text{Direct}_{s_j \to y4_{NRN}} = s_{j(y4_{NRN})}$$

$$\text{Indirect}(1)_{s_j \to y4_{NRN}} = \lambda_{24} s_{j(y2_{VV})}$$

$$\text{Total}_{s_j \to y1_Y} = \text{Direct}_{s_j \to y4_{NRN}} + \text{Indirect}(1)_{s_j \to y4_{NRN}} = s_{j(y4_{NRN})} + \lambda_{24} s_{j(y2_{VV})}$$

The Manhattan plots of direct (A), indirect (B), and total (C) SNP effects on NRN are presented in Fig. 6.

Vegetative vigor

The HC algorithm did not identify any mediator trait for vegetative vigor (VV) (Fig. 4). Therefore, the genetic architecture of VV was seemingly controlled only by the direct SNP effect. The total effect of the j th SNP on VV consists of its direct effect (Fig. 7).

$$\text{Direct}_{s_j \to y2_{VV}} = s_{j(y2_{VV})}$$

$$\text{Total}_{s_j \to y2_{VV}} = Direct_{s_{j(y2_{VV})}} = s_{j(y2_{VV})}$$

The Manhattan plots of direct (A) and total (B) SNP effects on VV are presented in Fig. 7.

Fruit Size (FS)

Similar to VV, the HC algorithm did not identify a mediator trait for fruit size (FS) (Fig. 4). Therefore, the SNP effect for FS is given by only its direct SNP effect.

$$\text{Direct}_{s_j \to y3_{FS}} = s_{j(y3_{FS})}$$

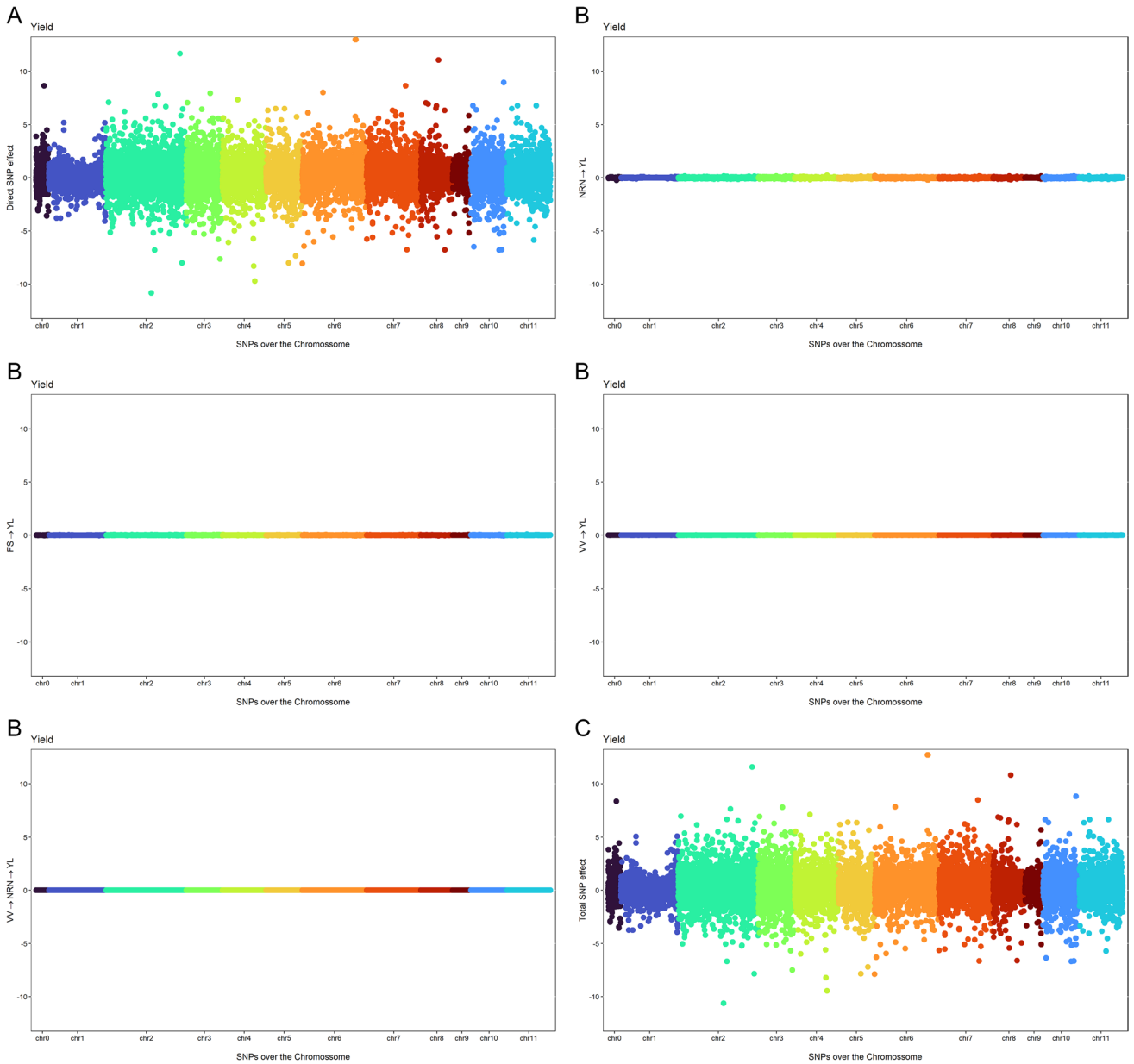

Fig. 5 Manhattan plots for direct (**A**), indirect (**B**), and total (**C**) single-nucleotide polymorphism effects on yield (YL) obtained using SEM-GWAS based on the network structure learned by the Hill Climbing algorithm. *VV* vegetative vigor, *NRN* number of reproductive nodes

$$\text{Total}_{s_j \to y3_{FS}} = \text{Direct}_{s_{j(y3_{FS})}} = s_{j(y3_{FS})}$$

The Manhattan plots of direct (A) and total (B) SNP effects on VV are presented in Fig. 8.

We observed that direct SNP effects were highly correlated ($R^2 > 0.98$) with total SNP effects for all traits. The indirect SNP and total SNP effects were positively correlated for VV → NNR (0.02) and VV → YL (0.03) and negatively correlated for NRN → YL (-0.72), FS → YL (−0.14) and VV → NNR → YL (−0.03).

Structural equation model genome-wide association study

A total of 87 SNP were statistically significant ($q < 0.01$) for NRN and YL. Among those, 39 and 48 SNP showed a significant association for NRN and YL, respectively (Table S1). In addition, we explored the functionalities of regions associated with these significant SNP (Table S1). It is important to emphasize that *C. arabica* is an allotetraploid from *C. canephora* and *Coffea eugenioides* (Lashermes et al. 1999). Therefore, the genome of *C. arabica* is divided into

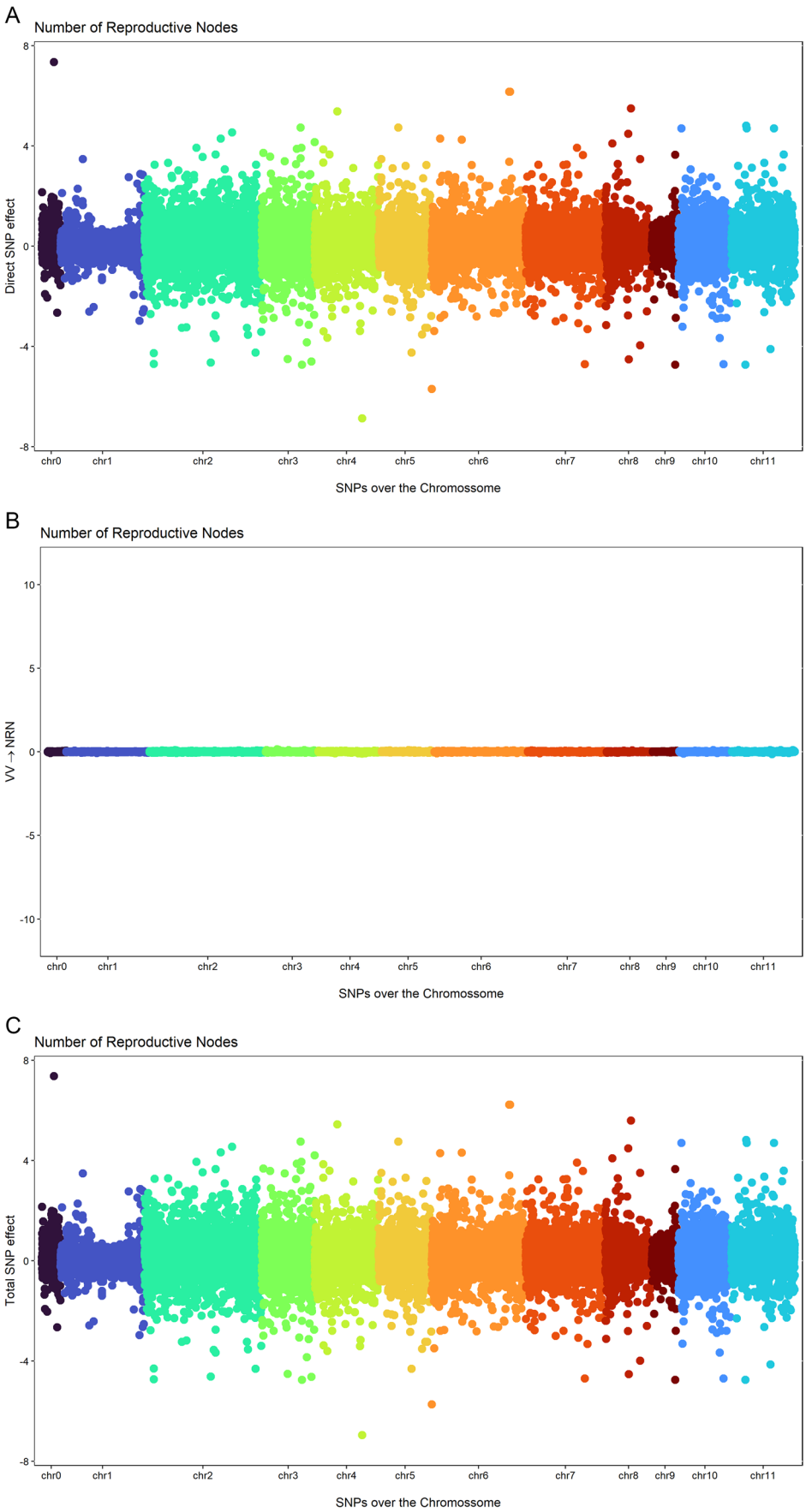

Fig. 6 Manhattan plots for direct (**A**), indirect (**B**), and total (**C**) SNP effects on the number of reproductive nodes (NRN) obtained using SEM-GWAS based on the network structure learned by the Hill Climbing algorithm. *VV* vegetative vigor

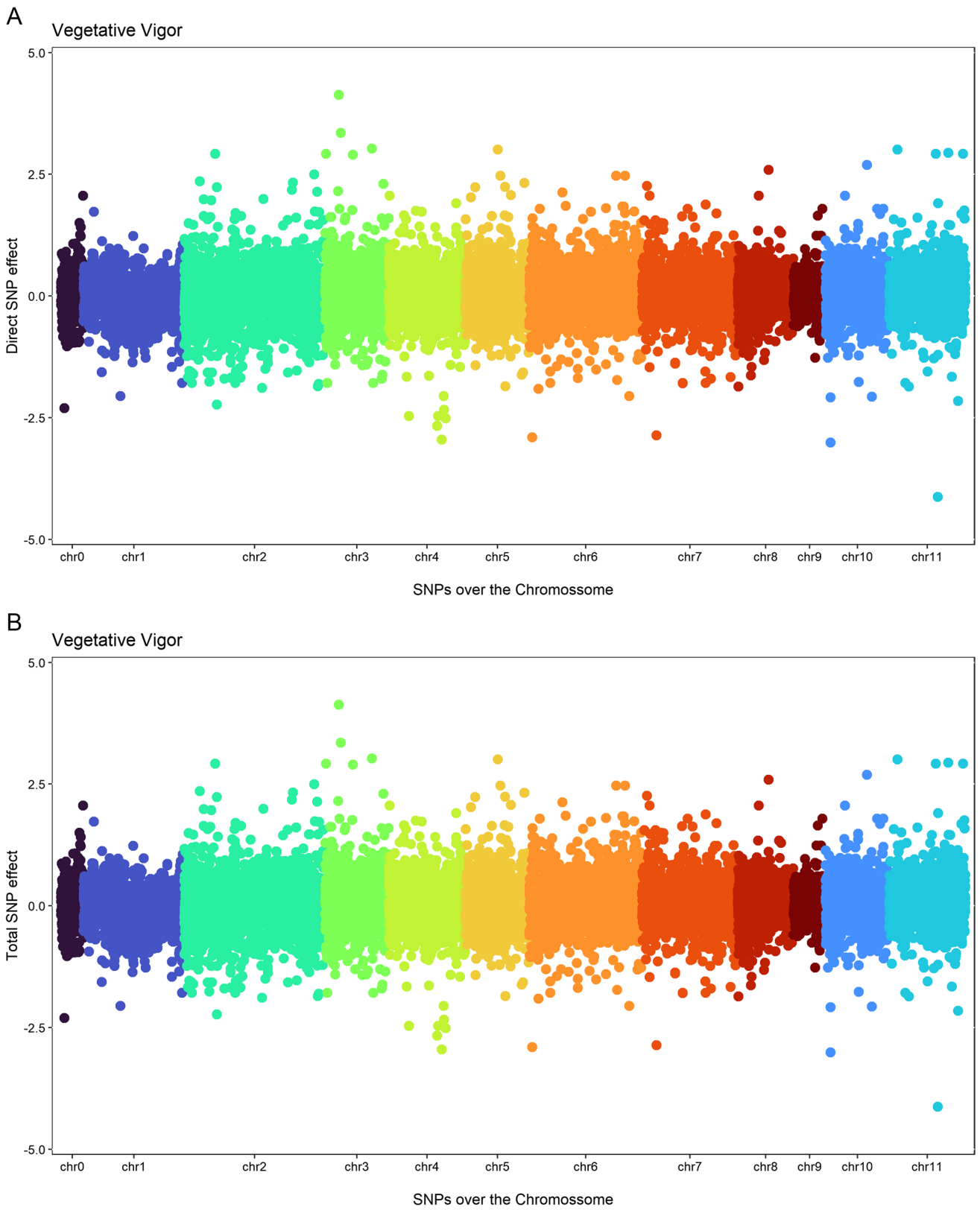

Fig. 7 Manhattan plots for direct (**A**) and total (**B**) SNP effects on vigor vegetative using SEM-GWAS-based model by including the network structure learned from the Hill Climbing algorithm

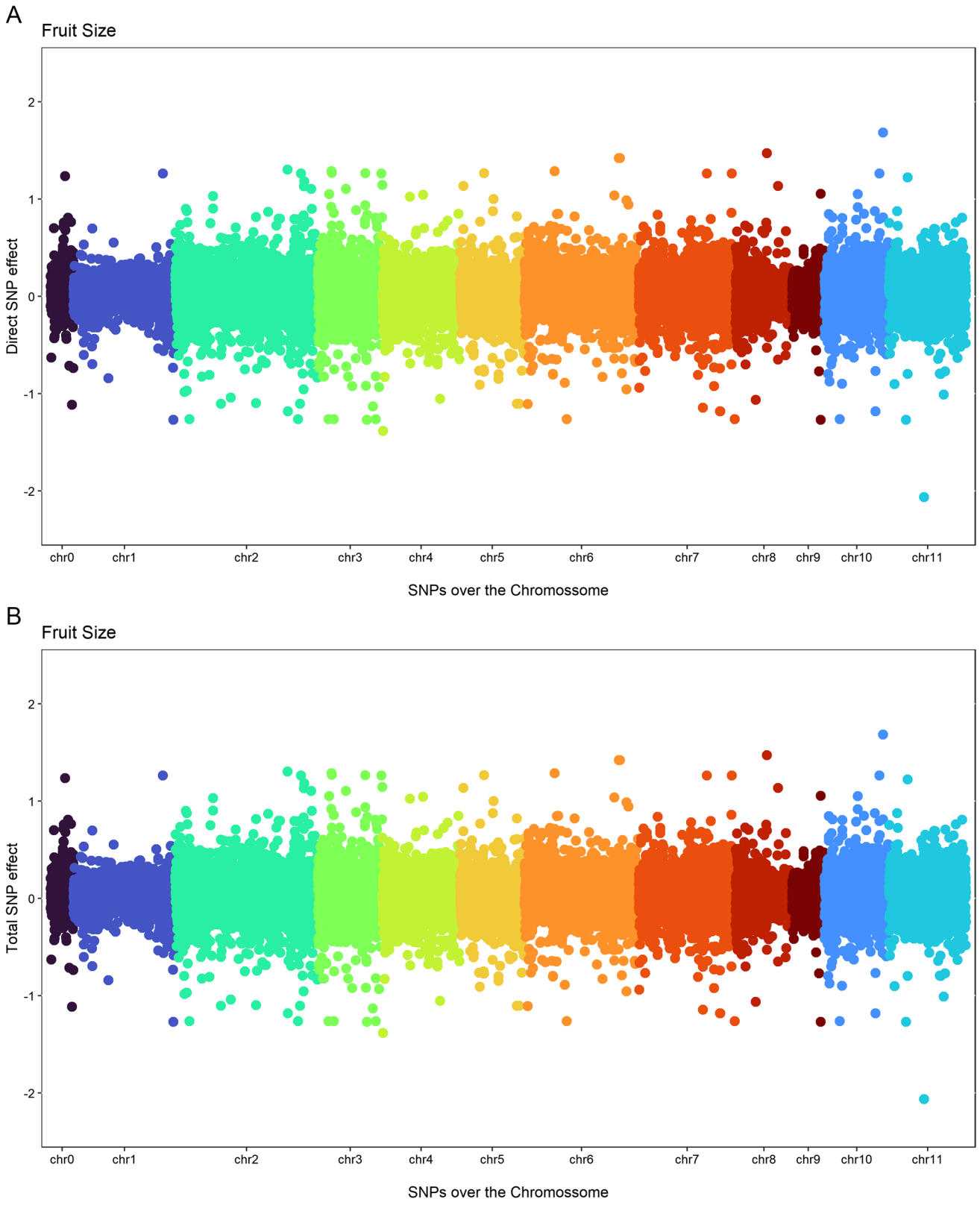

Fig. 8 Manhattan plots for direct (**A**) and total (**B**) SNP effects on fruit size using SEM-GWAS based model by including the network structure learned from the Hill Climbing algorithm

two subgenomes. For this reason, together with the information of each SNP marker, a code was added for subgenome information, "c" and "e" referring to *C. canephora* and *C. eugenioides*, respectively (Table S1). Although there are few genetic mapping studies in coffee, most previous studies were carried out in *C. canephora*. Moncada et al. (2016) found six QTL for yield on chromosomes 2, 4, and 11 using *C. arabica*.

Pathway enrichment analyses

Several ontologies and pathways were enriched (P value < 0.01) for the traits investigated (Figs. 9 and 10). For instance, pathways connected hydrolase activity (1 gene), metal ion binding (3 genes), isomerase activity (1 gene), transferase activity (2 genes), ADP binding (2 genes), ATP binding (4 genes), DNA binding (1 gene), catalytic activity (5 genes), ligase activity (1 gene), oxidoreductase activity (4 genes), phospholipid scramblase activity (1 gene), ubiquitin protein ligase binding (1 gene), and NAD^+ binding (1 gene) for Y. For NRN, we have pathways connected isomerase activity (1 gene), protein dimerization activity (1 gene), calmodulin binding (1 gene), ATP binding (1 gene), transferase activity (1 gene), metal ion binding (1 gene), and histone demethylase activity (1 gene).

For the YL KEGG analysis, we can observe that oxidative phosphorylation pathway, NFU1 iron-sulfur cluster protein pathway, ubiquitin-mediated proteolysis pathway,

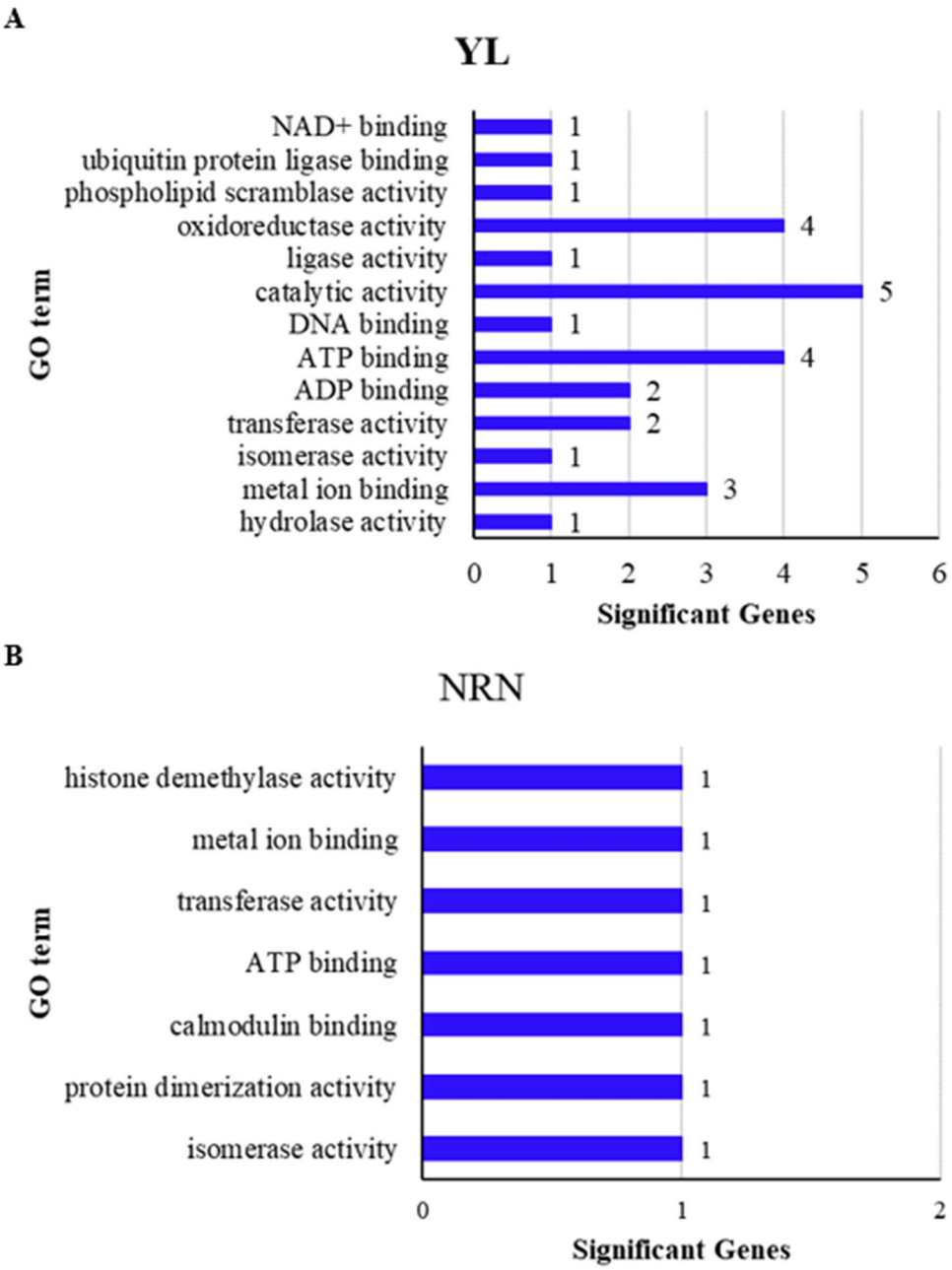

Fig. 9 Significantly enriched GO terms for the investigated traits. **A** Yield (YL); **B** number of reproductive nodes (NRN). SNPs obtained from MTM-GWAS (P value < 0.01)

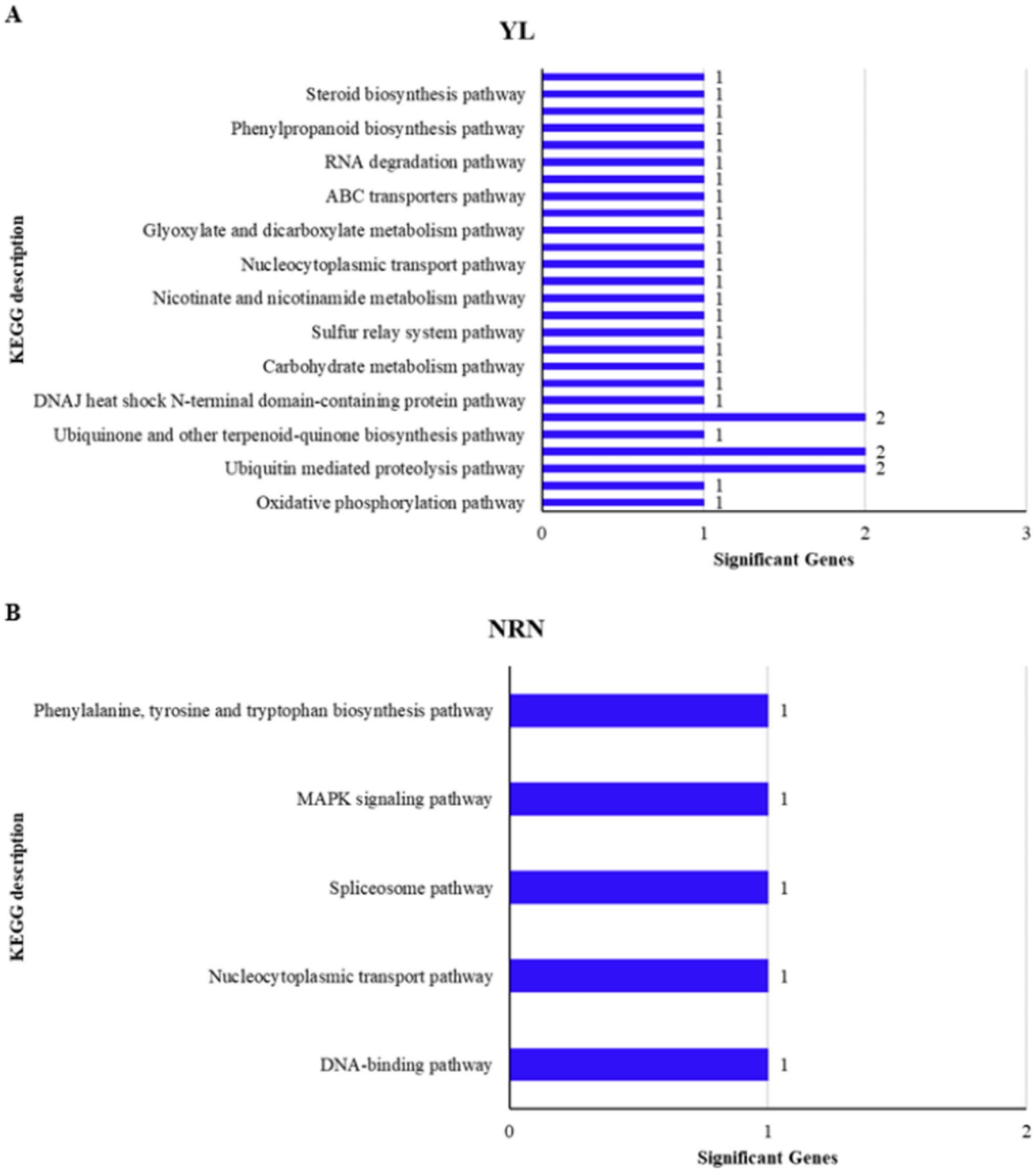

Fig. 10 Significantly enriched KEGG pathways for the investigated traits. **A** Yield (YL); **B** number of reproductive nodes (NRN). SNPs obtained from MTM-GWAS (P value < 0.01)

glycerophospholipid metabolism pathway, ubiquinone and other terpenoid-quinone biosynthesis pathway, spliceosome pathway, DNAJ heat shock N-terminal domain-containing protein pathway, putative receptor-like protein kinase At3g47110 pathway, carbohydrate metabolism pathway, phenylalanine, tyrosine, and tryptophan biosynthesis pathway, sulfur relay system pathway, plant-pathogen interaction pathway, nicotinate and nicotinamide metabolism pathway, AP endonuclease pathway, nucleocytoplasmic transport pathway, glycolysis/gluconeogenesis pathway, glyoxylate

and dicarboxylate metabolism pathway, thylakoid lumenal protein pathway, ABC transporters pathway, plant hormone signal transduction pathway, RNA degradation pathway, ubiquitin mediated proteolysis pathway, phenylpropanoid biosynthesis pathway, linoleic acid metabolism pathway, steroid biosynthesis pathway, and pentose phosphate pathway. For NRN KEGG analysis, we can observe that DNA-binding pathway, nucleocytoplasmic transport pathway, spliceosome pathway, MAPK signaling pathway and phenylalanine, tyrosine, and tryptophan biosynthesis pathway.

Discussion

Genetic parameters

The narrow sense heritability estimates for YL (0.14 ± 0.10) and VV (0.39 ± 0.14) were consistent with those reported in the literature. Specifically, the estimated heritability of YL was close to the estimates reported by Carvalho et al. (2019) and Bergo et al. (2020) based on 246 *C. canephora* genotypes (0.15) and 46 clones of *C. canephora* (0.16), respectively. For VV, our heritability estimate was close to those reported by Sousa et al. (2019) using 245 F_2 plants of *C. canephora* (0.34). For FS (0.61 ± 0.12) and NRN (0.13 ± 0.17), the heritability estimates differed from those in the literature, possibly due to the different method used and the population used. For example, Bikila and Sakiyama (2017) and Sousa et al. (2019) reported the FS heritability estimates of 0.42 and 0.36, respectively. Sousa et al. (2019) found the heritability estimate of 0.23 for NRN. Genetic correlation estimates were not significant for any of the traits. Our finding agrees with Bikila and Sakiyama (2017), who reported no significant genetic correlation between YL and FS. Even though there were no significant genetic correlations, the multivariate methodology was used, as it is unprecedented in the culture of *Coffea arabica*.

Structural equation model genome-wide association study

This study aimed to understand the genetic interdependence of evaluated traits (YL, VV, FS, and NRN) in *Coffea arabica* using the structural equations modeling genome-wide association study (SEM-GWAS) approach. According to Momen et al. (2019), SEM-GWAS can help to better understand the underlying biological mechanisms by distinguishing the source of SNP effects into direct and indirect effects. Additionally, Pegolo et al. (2020) emphasized that SEM-GWAS offers a powerful and flexible approach to capture interrelated structures missed through MTM-GWAS. In this study, the Bayesian network-aided SEM-GWAS allowed us to obtain the interrelated network and their path coefficients and to partition the total SNP effects. We observed the positive direct (VV → YL) and the indirect (VV → NNR → YL) association between VV and YL, two important economic traits in Arabica coffee. The positive interrelationship between VV and YL indicates that the better the vegetative vigor status of the plant, the greater its production will be.

We used SEM-GWAS to decompose pleiotropic QTL into direct and indirect effects. The direct SNP effect was higher than the indirect SNP effect for YL and NRN (Figs. 5 and 6), indicating that the total SNP effects from YL and NRN are driven largely by genetic effects acting directly on them rather than indirect effects. We did not find a QTL that controls the traits together. In contrast, Momen et al. (2019) reported that the same QTL in rice controlled projected shoot area and water use efficiency. We found associations of 87 SNP for YL, NRN, FS, and VV, and the positional candidate genes for these traits around significant SNP were investigated.

Several candidate genes for YL were found based on the functional annotation. For example, (i) LOC113731461_e–Stress enhanced protein 1, Chloroplastic; (ii) LOC113714295_c–Abscisic stress-ripening protein 5; and (iii) LOC113726102_c–negative regulator of systemic acquired resistance (SAR) SNI1. According to Heddad and Adamska (2000), the Stress enhanced protein 1, chloroplastic may play a photoprotective role in the thylakoid membrane in response to light stress in *Arabidopsis thaliana*, thus providing greater photosynthetic efficiency and increasing the production. Chloroplastic may play a photoprotective role in the thylakoid membrane response to light stress. Li et al. (2017) identified the involvement of Abscisic stress-ripening protein 5 in drought tolerance in rice, potentially playing a positive role in response to water stress, regulating abscisic acid (ABA) biosynthesis, promoting stomatal closure, and acting as a protein similar to chaperone that possibly prevents the inactivation of proteins related to water stress. Durrant et al. (2007) reported a negative reduction in pathogenesis-related protein expression and DNA recombination during susceptible pathogen infection in SAR in *Arabidopsis*. Therefore, SNI1 is involved in short-term defense response and a long-term supply strategy. However, no mechanism was identified that directly influences NRN control.

In general, our findings indicate that using SEM-GWAS for analyzing a set of genetically related traits (YL, NRN, FS, and VV) in *Coffea arabica* resulted in a better understanding of the genetic interdependence of those traits. The evaluated four traits are mostly driven by genetic effects acting directly on those traits. The use of MTM-GWAS does not incorporate the interdependence between the traits since it only expresses overall genetic effects. In terms of breeding, knowledge of the interrelationships between traits can help design the best selection strategy, where high direct effects associated with a high correlation indicate that the trait in question is the main

determinant of variations in the other variable, thus waiting that indirect selection is effective, if the opposite occurs (high direct effect and low correlation), it may indicate that the trait should not be completely discarded from use in indirect selections, as simultaneous selection may provide good results.

This study represents the first application of SEM-GWAS in coffee. In conclusion, SEM-GWAS decomposes the genetic interrelationships between four arabica coffee traits (YL, NRN, FS, and VV). Among those traits, only YL and NRN showed indirect effects. We detected significant genomic regions and identified candidate genes that act directly on YL. There was no evidence of a QTL controlling the same traits jointly.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11295-023-01597-8.

Acknowledgements This work was financially supported by the Brazilian Coffee Research and Development Consortium (Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café–CBP&D/Café), by the Foundation for Research Support of the state of Minas Gerais (FAPEMIG), by the National Council of Scientific and Technological Development (CNPq), and by the National Institutes of Science and Technology of Coffee (INCT/Café). MMS, MN, and CFA is supported by a master's scholarships (140877/2021-5) and scientific productivity (307798/2019-4 and 306772/2020-5), respectively, from Brazilian Council for Scientific and Technological Development (CNPq).

Author contribution MMS designed the study, analyzed the data, and wrote the first draft of the paper. CFA and ACCN revised the paper. ETC provided phenotypic data from the breeding program and revised the paper. ACBO provided phenotypic data from the breeding program. MM and GM supported the data analysis and revised the paper. MN designed the study, supported the data analysis, and revised the paper. GM and MN supervised the study.

Funding Not applicable.

Data availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no conflict of interest.

Data archiving statement The authors have not submitted biological data to any of the public databases.

References

Barka GD, Caixeta ET, de Almeida RF, Alvarenga SM, Zambolim L (2017) Differential expression of molecular rust resistance components have distinctive profiles in Coffea arabica-*Hemileia vastatrix* interactions. Eur J Plant Pathol 149:543–561. https://doi.org/10.1007/s10658-017-1202-0

Bergo CL, Miqueloni DP, Lunz AMP et al (2020) Estimation of genetic parameters and selection of *Coffea canephora* progenies evaluated in Brazilian Western Amazon. Embrapa, Acre

Bikila BA, Sakiyama NS (2017) Estimation of genetic parameters in *Coffea canephora* Var. Robusta Adv Crop Sci Technol 5:310

Borém A, Miranda GV, Fritsche-Neto R (2021) Melhoramento de plantas. Oficina de Textos

Carvalho HF, Silva FLD, Resende MDVD et al (2019) Selection and genetic parameters for interpopulation hybrids between kouilou and robusta coffee. Bragantia 78:52–59. https://doi.org/10.1590/1678-4499.2018124

Cilas C, Bar-Hen A, Montagnon C et al (2006) Definition of architectural ideotypes for good yield capacity in *Coffea canephora*. Ann Bot 97:405–411. https://doi.org/10.1093/aob/mcj053

DCCC (2019) Coffee market China is growing larger–coffee consumption and imports China. https://www.dccchina.org/2019/09/coffee-market-china-is-growing-larger-coffee-consumption-imports-opportunities-in-china/ Accessed 15 March 2022

de Los Campos G, Hickey JM, Pong-Wong R et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. https://doi.org/10.1534/genetics.112.147983

Department of Agriculture. https://apps.fas.usda.gov/psdonline/circulars/coffee.pdf. Accessed 29 July 2022

Durrant WE, Wang S, Dong X (2007) Arabidopsis SNI1 and RAD51D regulate both gene transcription and DNA recombination during the defense response. Proc Natl Acad Sci 104:4223–4227. https://doi.org/10.1073/pnas.0609357104

ESTADOS UNIDOS (2021) Coffee: world markets and trade. USDA.gov - United States

Ferrão RG, da Fonseca AFA, Ferrão MAG et al (2012) Café Conilon: Técnicas de Producion com variedades mejoradas. Incaper, Vitória

Ferrão RG, de Muner LH, da Fonseca AFA et al (2016) Café Conilon. Incaper, Vitória

Gianola D, Sorensen D (2004) Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics 167:1407–1424. https://doi.org/10.1534/genetics.103.025734

Gimase JM, Thagana WM, Omondi CO et al (2020) Genome-wide association study identify the genetic loci conferring resistance to Coffee Berry disease (*Colletotrichum kahawae*) in *Coffea arabica* var. Rume Sudan Euphytica 216:1–17. https://doi.org/10.1007/s10681-020-02621-x

Heddad M, Adamska I (2000) Light stress-regulated two-helix proteins in Arabidopsis thaliana related to the chlorophyll a/b-binding gene family. Proc Natl Acad Sci 97:3741–3746. https://doi.org/10.1073/pnas.97.7.3741

Korb KB, Nicholson AE (2010) Bayesian artificial intelligence. CRC Press, Florida

Korte A, Vilhjálmsson BJ, Segura V et al (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44:1066–1071. https://doi.org/10.1038/ng.2376

Lashermes P, Combes MC, Robert J et al (1999) Molecular characterisation and origin of the *Coffea arabica* L. genome. Mol Genet Genom 261:259–266. https://doi.org/10.1007/s004380050965

Li J, Li YL, Yin Z et al (2017) OsASR5 enhances drought tolerance through a stomatal closure pathway associated with ABA and H2 O2 signalling in rice. Plant Biotechnol J 15:183–196. https://doi.org/10.1111/pbi.12601

Meyer K (2007) WOMBAT—a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML).

J Zhejiang Univ Sci B 8:815–821. https://doi.org/10.1631/jzus.2007.B0815
Meyer K, Tier B (2012) "SNP Snappy": a strategy for fast genome-wide association studies fitting a full mixed model. Genetics 190:275–277. https://doi.org/10.1534/genetics.111.134841
Mishra MK, Slater A (2012) Recent advances in the genetic transformation of coffee. Biotechnol Res Int 2012:1–17. https://doi.org/10.1155/2012/580857
Momen M, Mehrgardi AA, Roudbar MA et al (2018) Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models. Front Genet 9:455–466. https://doi.org/10.3389/fgene.2018.00455
Momen M, Campbell MT, Walia H et al (2019) Utilizing trait networks and structural equation models as tools to interpret multi-trait genome-wide association studies. Plant Methods 15:1–14. https://doi.org/10.1186/s13007-019-0493-x
Moncada MDP, Tovar E, Montoya JC et al (2016) A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. Tree Genet Genomes 12:1–17. https://doi.org/10.1007/s11295-015-0927-1
Nonato JVA, Carvalho HF, Borges KLR et al (2021) Association mapping reveals genomic regions associated with bienniality and resistance to biotic stresses in arabica coffee. Euphytica 217:1–19. https://doi.org/10.1007/s10681-021-02922-9
O'Reilly PF, Hoggart CJ, Pomyen YL et al (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One 7:e34861
Pegolo S, Momen M, Morota G et al (2020) Structural equation modeling for investigating multi-trait genetic architecture of udder health in dairy cattle. Sci Rep 10:1–15. https://doi.org/10.1038/s41598-020-64575-3
Resende MDVD (2016) Software Selegen-REML/BLUP: a useful tool for plant breeding. Crop Breed Appl Biotechnol 16:330–339
Romero G, Vásquez LM, Lashermes P et al (2014) Identification of a major QTL for adult plant resistance to coffee leaf rust (*Hemileia vastatrix*) in the natural Timor hybrid (*Coffea arabica* x *C. canephora*). Plant Breed 133:121–129. https://doi.org/10.1111/pbr.12127
Sant'Anna GC, Pereira LF, Pot D et al (2018) Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica* L. Sci Rep 8:1–12. https://doi.org/10.1038/s41598-017-18800-1
Scutari M (2010) Learning Bayesian Networks with the bnlearn R Package. J Stat Softw 35:1–22. https://doi.org/10.48550/arXiv.0908.3817
Scutari M, Denis JB (2014) Bayesian networks: with examples in R. Chapman and Hall, CRC, New York
Shim H, Chasman DI, Smith JD et al (2015) A multivariate genome-wide association analysis of 10 LDL Subfractions, and Their Response to Stati Treatment, in 1868 Caucasians. PLoS ONE 10:1–20. https://doi.org/10.1371/journal.pone.0120758
Sousa TV, Caixeta ET, Alkimim ER, de Oliveira ACB, Pereira AA, Sakiyama NS et al (2017) Population structure and genetic diversity of coffee progenies derived from Catuaí and Híbrido de Timor revealed by genome-wide SNP marker. Tree Genet Genomes 13:124. https://doi.org/10.1007/s11295-017-1208-y
Sousa TV, Caixeta ET, Alkimim ER et al (2019) Early selection enabled by the implementation of genomic selection in *Coffea arabica* breeding. Front Plant Sci 9:1934–1946. https://doi.org/10.3389/fpls.2018.01934
Sousa ICD, Nascimento M, Silva GN et al (2020) Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Sci Agric 78:1–8. https://doi.org/10.1590/1678-992X-2020-0021
Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci 100:9440–9445. https://doi.org/10.1073/pnas.153050910
Tran HT, Furtado A, Vargas CAC et al (2018) SNP in the *Coffea arabica* genome associated with coffee quality. Tree Genet Genomes 14:1–15. https://doi.org/10.1007/s11295-018-1282-9
Valente BD, Rosa GJ, Gianola D, Wu X-L, Weigel KA (2013) Is structural equation modeling advantageous for the genetic improvement of multiple traits? Genetics 194:561–572. https://doi.org/10.1534/genetics.113.151209
VanRaden PM (2008) Efficient methods to compute genomic predictions. Int J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980
Wallace JG, Rodgers-Melnick E, Buckler ES (2018) On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. Annu Rev Genet 52:421–444. https://doi.org/10.1146/annurev-genet-120116-024846
Wang Z, Chapman D, Morota G et al (2020) A multiple-trait Bayesian variable selection regression method for integrating phenotypic causal networks in genome-wide association studies. G3: Genes Genomes Genet 10:4439–4448. https://doi.org/10.1534/g3.120.401618
Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702
Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44:821–824. https://doi.org/10.1038/ng.2310