UNIVERSIDADE FEDERAL DE VIÇOSA

MATHEUS MASSARIOL SUELA

STRUCTURAL EQUATION MODELS FOR GENOME-WIDE ASSOCIATION STUDY IN *Coffea arabica*

VIÇOSA - MINAS GERAIS
2021

MATHEUS MASSARIOL SUELA

STRUCTURAL EQUATION MODELS FOR GENOME-WIDE ASSOCIATION STUDY IN *Coffea arabica*

Dissertation submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

Adviser: Moysés Nascimento

Co-advisers: Camila Ferreira Azevedo
Eveline Teixeira Caixeta Moura
Gota Morota

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
S944s
2021

Suela, Matheus Massariol, 1997-
Structural equation models for genome-wide association study in *Coffea arabica* / Matheus Massariol Suela. – Viçosa, MG, 2021.
1 dissertação eletrônica (59 f.): il. (algumas color.).

Inclui anexos.
Orientador: Moysés Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2021.
Referências bibliográficas: f. 48-50.
DOI: https://doi.org/10.47328/ufvbbt.2021.164
Modo de acesso: World Wide Web.

1. Café. 2. Modelagem de equações estruturais. 3. Marcadores genéticos. 4. Genômica. I. Nascimento, Moysés. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 633.73

Bibliotecário(a) responsável: Renata de Fátima Alves x

MATHEUS MASSARIOL SUELA

STRUCTURAL EQUATION MODELS FOR GENOME-WIDE ASSOCIATION STUDY IN *Coffea arabica*

Dissertation submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

APPROVED: July 27, 2021

Assent:

Matheus Massariol Suela
Author

Moysés Nascimento
Adviser

Aos meus pais, Geraldo e Rosimeri.

DEDICO.

AGRADECIMENTOS

A toda minha família, pelo imenso apoio durante minha vida, em especial ao meu pai Geraldo Guerino Suela, que sempre tira um instante de seu tempo para me enriquecer com seus ensinamentos técnicos e de vida e a minha mãe Rosimeri Massariol Suela, que sempre depositou em mim seu amor incondicional.

A minha namorada Letícia, pelo o amor, carinho, preocupação e confiança durante toda nossa jornada, ora nos momentos difíceis, ora nos momentos de vitórias, tornando minha caminhada muito mais segura e alegre.

Ao meu orientador, Moysés Nascimento, pela paciência, apoio e pela grande amizade.

A minha coorientadora, Camila Ferreira Azevedo, que desde a graduação me enriqueceu com suas experiências e com sua amizade.

Aos meus coorientadores Gota Morota (*Virginia Polytechnic Institute and State University*) e Eveline Teixeira Caixeta Moura (Embrapa – Café), pela ajuda em análises e disponibilização de dados.

Ao amigo, Mehdi Momen (*University of Wisconsin* – Madinson), que me ajudou em análises chaves deste projeto.

Aos meus amigos do LICAE (Laboratório de Inteligência Computacional e Aprendizado Estatístico), que foram minha família, tornando minha jornada muito mais alegre e descontraída em Viçosa.

À banca, composta por Ana Carolina Nascimento e Camila Ferreira Azevedo que aceitaram o convite que lhes foi feito e, dessa forma, colaboraram para a conclusão deste projeto.

À agência de fomento CNPq, que colaborou para que essa pesquisa fosse possível.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

BIOGRAFIA

MATHEUS MASSARIOL SUELA, filho de Rosimeri Massariol Suela e Geraldo Guerino Suela, nasceu em Colatina, Espírito Santo, em 22 de março de 1997.

Em março de 2012, ingressou no curso técnico em agropecuária no Instituto Federal do Espírito Santo – Campus Itapina, formando-se em dezembro de 2014.

Ingressou no curso de Agronomia, em março de 2015, na Universidade Federal de Viçosa, Minas Gerais – MG, graduando-se em janeiro de 2020.

Em março do mesmo ano, iniciou o curso de mestrado do Programa de Pós-Graduação em Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 26 de julho de 2021.

RESUMO

SUELA, Matheus Massariol, M.Sc., Universidade Federal de Viçosa, julho de 2021. **Structural equation models for genome-wide association study in *Coffea arabica***. Orientador: Moysés Nascimento. Coorientadores: Camila Ferreira Azevedo, Eveline Teixeira Caixeta e Gota Morota.

O melhoramento em café foi baseado em técnicas clássicas por muito tempo, porém, com o advento de técnicas genômicas e de fenotipagem de precisão, os programas de melhoramento vêm apresentando melhores resultados e mais velozes, mesmo com os programas se tornando cada vez mais complexos, em termos de quantidades e tipos de características estudadas. Dessa forma, a existência de interrelações entre caracteres podem gerar impactos importantes em um programa de melhoramento, como por exemplo, na descoberta de regiões genômicas que contribuem para determinadas características. Especificamente, tais características podem atuar tanto de forma direta quanto indireta na característica em estudo. Sabendo disso, compreender os efeitos diretos e indiretos que um caráter exerce em outro, é de grande importância para a fase de seleção. Tradicionalmente, para realizar o estudo das associações entre características, técnicas multivariadas são aplicadas, porém, são tais metodologias negligenciam as inter-relações entre as mesmas. Dessa forma, a utilização da Rede Bayesiana (BN) em conjunto com Modelo de Equações Estruturadas (SEM) sob o enfoque do estudo de associação genômica ampla (GWAS), permite quantificar o efeito dos marcadores, particionando seus valores em efeitos diretos e indiretos para as características presentes na rede formada. Com o objetivo de explorar estas inter-relações, foram analisados fenótipos relacionados às características morfológicas (tamanho do fruto, número de nós reprodutivos), fisiológicas (vigor vegetativo) e produtivas (produção) em 195 genótipos de *Coffea arabica*, provenientes de uma parceria entre a Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG), Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e Universidade Federal de Viçosa (UFV). A rede fenotípica inferida por meio do algoritmo Hill Climbing foi usada para estimar os coeficientes estruturais. Realizando uma integração entre modelos multivariados - GWAS e SEM-GWAS foi possível identificar inter-relação positiva entre vigor vegetativo em produção e de vigor vegetativo pra número de nós reprodutivos e negativo de número de nós reprodutivos e tamanho do fruto para produção. Também foi possível detectar regiões genômicas significativas, e assim, identificar três genes que atuam diretamente sobre produção.

Palavras-chave: *Coffea arabica*. Redes Bayesianas. Modelo de Equações Estruturadas. GWAS.

ABSTRACT

SUELA, Matheus Massariol, M.Sc., Universidade Federal de Viçosa, July, 2021. **Structural equation models for genome-wide association study in *Coffea arabica***. Adviser: Moysés Nascimento. Co-advisers: Camila Ferreira Azevedo, Eveline Teixeira Caixeta and Gota Morota.

Coffee breeding techniques were based on classical techniques for a long time, however, with the advent of genomic techniques and precision phenotyping, breeding programs have been showing best and faster results, even with the programs becoming more complex, in terms of quantities and types of characteristics studied. Thus, the existence of interrelationships between characters can generate important impacts in a breeding program, such as the discovery of genomic regions that contribute to certain characteristics, these can act directly, or indirectly. Knowing this, understanding the direct and indirect effects that one character has on another is of great importance for the selection phase. Traditionally, multivariate techniques are applied, but phenotypic interrelationships are neglected. Thus, the use of the Bayesian Network (BN) in conjunction with the Structured Equation Model (SEM) under the focus of the genomic wide association study (GWAS), allows quantifying genetic parameters, partitioning such values into direct and indirect effects for the traits. present in the formed network. In order to explore these interrelationships, they were able to phenotypes related to morphological (fruit size and number of reproductive nodes), physiological (vegetative vigor) and productive (production) characteristics in 195 *Coffea arabica* genotypes from a partnership between Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG), Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) and Federal University of Viçosa (UFV). The phenotypic network inferred by means of the Hill Climbing algorithm was used to estimate the appropriate coefficients. By performing an integration between multivariate models - GWAS and SEM-GWAS it was possible to identify a positive interrelationship between vegetative vigor in yield and vegetative vigor for the number of reproductive nodes and negative for the number of reproductive nodes and fruit size for yield. It was also possible to detect significant genomic regions, and thus identify three genes that act directly on yield.

Keywords: *Coffea arabica*. Bayesian Network. Structural Equation Models. GWAS.

SUMMARY

GENERAL INTRODUCTION

Coffee is the second most important commodity in international trade, after crude oil (MISHRA, 2019), and has a very large impact in several countries in Asia, Africa and Latin America, both in economic and social terms. The total world production in the 2020/21 harvest is equivalent to approximately 10.5 million tons processed (USDA, 2021). Brazil is the world is largest producer and exporter of coffee (arabica and conilon), with a total of approximately 3.8 million tons of processed coffee produced in the 2020/21 harvest, according to data from the Companhia Nacional de Abastecimento (CONAB, 2021). This production comes from approximately 2.2 million hectares, being disposed in around 277.3 thousand hectares in the formation phase and around 1.89 million hectares in the production phase (CONAB, 2021). Of the total produced in the 2020/21 harvest, approximately 77.3% comes from arabica coffee, the other 22.7% comes from conilon coffee.

The genus *Coffea* belongs to the Rubiaceae family and consists of more than 125 species (DAVIS, 2011; DAVIS et al. 2006; RAZAFINARIVO et al. 2013). However, commercially, there are two species that stand out, the arabica coffee (*Coffea arabica*) and the conilon coffee (*Coffea canephora*). According to Ferrão et al. (2017) and Carvalho (1946) the species *C. canephora* differs from *C. arabica* in several agronomic characteristics, which from the viewpoint of genetic improvement are very important, namely: i) it has a multi-stemmed shrub; ii) larger leaves, well wavy, with a lighter green coloration; iii) self-incompatible flowers; iv) fruits a little more spherical, smaller, with red, yellow and orange color when ripe and thinner exocarp; v) seeds of variable size, with a well-adhering silvery skin, green endosperm and higher caffeine content. When it comes to the genome, *C. arabica* is a tetraploid plant (2n = 4X = 44), while *C. canephora* is a diploid (2n = 2X = 22).

According to (MISHRA, 2019), the genetic improvement programs for coffee, initially with Arabica and only from the 1950s onwards with conilon, initially aimed at increasing productivity and resistance to rust, only from 1990 onwards than others characteristics, such as beverage quality, pest and drought resistance gained notoriety. This start was made using conventional breeding techniques, but this became a major bottleneck, as from the selection of parents, through hybridization until finally reaching the progeny evaluations, approximately 30 years are required to develop a new cultivar, in addition to which becomes quite an expensive process. Thus, several strategies had to be implemented for the gains to be greater. One of the applied strategies was the genomic association, which consists in the application of

methodologies in order to detect significant markers for certain characteristics of interest. This technique allows the quantification of the effects of markers on the evaluated trait, but currently, breeding programs have been using several traits at once, since certain traits can positively, negatively or not affect one another. Thus, this work proposes a new way of using GWAS (Genome-Wide Association Study) in morphological, pest and productive characteristics of Arabica coffee, using Bayesian Networks (BN) and Structured Equation Models (SEM), partition the effect of the marker into direct and indirect, allowing the analysis of the direct and indirect impact on the target trait compared to several others and also identify significant markers that represent a candidate gene.

REFERENCES

CONAB - ACOMPANHAMENTO DA SAFRA BRASILEIRA DE CAFÉ. Terceiro levantamento, Setembro de 2021. v. 8 - Safra 2021, n. 3.

MISHRA, M. K. 2019. Genetic resources and breeding of coffee (*Coffea* spp.). In: Advances in Plant Breeding Strategies: Nut and Beverage Crops. Springer, Cham, p. 475-515.

DAVIS, A. P.; GOVAERTS, R.; BRIDSON, D. M.; STOFFELEN, P. An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). Botanical Journal of the Linnean Society, London, v. 152, p. 465-512, July 2006.

DAVIS, A. P.; TOSH, J.; RUCH, N.; FAY, M. F. Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data, implications for the size, morphology, distribution and evolutionary history of Coffea. Botanical
Journal of the Linnean Society, London, v. 167, p. 1-21, Dec. 2011.

CARVALHO, A. Distribuição geográfica e classificação botânica do gênero Coffea com referência especial à espécie Arabica. Separata dos boletins da superintendência de serviços de café. Campinas, SP: IAC, dez. 1945 a abr. 1946.

Ferrão, R.G., de Muner, L.H., da Fonseca, A.F.A. and Ferrão, M.A.G., 2016. *Café Conilon*. Vitória, ES: Incaper, 2017.

Structural Equation Models for Genome-Wide Association Study in *Coffea arabica*

ABSTRACT

Yield is one of the most important characteristics for arabica coffee, however, it is affected by several other characteristics, even so, plant breeders search for to maximize this characteristic directly and/or indirectly, using characteristics that are often correlated. Thus, structural equation modeling (SEM) - GWAS was applied in order to explore interrelated dependencies between phenotypes related to morphology (fruit size, number of reproductive nodes), physiology (vegetative vigor) and productive (yield) characteristics in 195 *Coffea arabica* genotypes from a partnership between Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG), Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) and Universidade Federal de Viçosa (UFV). The phenotypic network inferred by means of the Hill Climbing algorithm was used to estimate the appropriate coefficients. By performing an integration between multivariate models - GWAS and SEM-GWAS it was possible to identify a positive interrelationship between vegetative vigor in production and vegetative vigor for the number of reproductive nodes and negative for the number of reproductive nodes and size of the fruit for production. It was also possible to detect significant genomic regions, and thus identify three genes that act directly on yield.

1 INTRODUCTION

Coffee is one of the most widely consumed beverages worldwide, with Brazil being the world's largest producer. Of all the coffee produced in the world (*Coffea canephora* and *Coffea arabica*), Brazil produces 39.76%, if we consider only Arabica coffee, the target of our study, this number already rises to 48.68% (USDA, 2021).

Due to the increase in coffee consumption in countries that were not as traditional, such as China (Nam, Z., 2014; DCCC, 2019), it is necessary to promote research that contributes to greater productivity and sustainability of production chain. In this context, genetic breeding is one of those responsible for promoting such advances in the midst of the development of cultures that meet the demands of the market (Oliveira et al., 2010; Carvalho et al., 2011; Barka et al., 2017). However, the improvement process takes time, since this culture has a long cycle, high size and a long juvenile period (Ferrão et al., 2017). Thus, it is recommended to apply innovative tools, such as the use of biotechnology, which can contribute to the genetic progress of the culture (Mishra and Slater, 2012; Ferrão et al., 2015).

Among these methodologies, genome-wide association studies (GWAS) have become increasingly popular for the elucidation of the genetic architecture of economically important traits (Momen et al., 2019). In coffee, GWAS have been successful in identifying regions on genome associated with a important of phenotypes, as example, yield, abiotic and biotic stresses, and plant morphological traits (Sant' Ana et al., 2018; Tran et al., 2018).

In breeding programs, correlated traits are recorded on the same material and the association mapping is performed independently for each trait. This approach can fail to study the genetic interdependence among traits and impose limitations on elucidating the genetic mechanisms underlying a complex system of traits (Momen et al., 2019). To circumvent this issue, the multi-trait GWAS (MTM-GWAS) was proposed. According to Zhou and Stephens (2012), Korte et al., (2012), O'reilly et al., (2012) and Momen et al. (2018) this approach reduces false positives and increases the statistical power of association tests in GWAS. Although MTM-GWAS is a valuable approach, this methodology does not inform how the traits are interrelated, that is does not provide information about causal relationships.

Momen et al. (2018) proposed to use Structural equation modeling for association studies (SEM-GWAS). According with these authors, compared to MTM-GWAS, the SEM-GWAS approach captures complex relationships and delivers a more comprehensive understanding of single nucleotide polymorphism (SNP) effects. Specifically, it can partition the total SNP

effects acting on a trait into direct and indirect effects enhancing our understanding of complex relationships among agronomic traits.

In a coffee breeding program context, some traits have an important impact on the culture. Among than, the Yield (Y), Vegetative Vigor (VV), Fruit Size (FS) and Number of Reproductive Nodes (NRN) deserves attention. According to Cilas et al. (2006) individuals who have larger amounts of NRN tend to have higher productions. According Ferrão et al., 2012, the FS is one of the main trait used to select production. The VV which shows your growth potential. Finally, the main trait in a breeding program is Yield (Y), which is, extremely impacted by several other characteristics at once.

In this context, we aimed to (1) estimate genetic parameters for phenological traits in the *Coffea arabica*; and (2) to enhance the understanding of the genetic architecture of these traits using SEM-GWAS approach.

2 MATERIAL AND METHODS

2.1 Phenotypic and Genotypic Data

The phenotypic and genotypic data comes from the *C. arabica* breeding program of the partnership between EPAMIG, UFV and EMBRAPA. An experimental area is maintained at the Department of Phytopathology – UFV (lat. 20°44′25" S, long. 42°50′52" W). This database contains 13 progenies from crosses between three parents of the Catuaí cultivar and three parents of the Hybrid of Timor (HdT), built in relation to coffee rust (*Hemileia vastatrix*). Fifteen genotypes were selected from each progenies mentioned above, totaling 195 individuals, which were genotyped for 21,211 SNP markers.

The genotypes were planted on February 11, 2011, using the spacing of 3.0 meters between rows and 0.7 meters between plants. Nutritional management was carried out following the requirements of the crop. More details can be seen in Sousa et al. (2019).

The phenotypic database used comprised four traits, which are: Yield (Y), Vegetative Vigor (VV), Number of Reproductive Nodes (NRN) and Fruit Size (FS). There was correction of the phenotypes for the effect of years, plots and years x plots interaction. The analyzes were performed considering the mixed linear models (REML/BLUP procedure), using the Selegen-REML/BLUP software (Resende, 2016b), using the following statistical model:

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

Where, $\mathbf{y}$ is the vector of data, $\mathbf{u}$ is the vector of general average in each year of evaluation, $\mathbf{g}$ is the vector of progeny effects, $\mathbf{p}$ is the vector of permanent variance between individuals, $\mathbf{r}$ is the vector of variance between types of populations, $\mathbf{b}$ is the vector of variance between plots, $\mathbf{i}$ is the vector of variance of the progenies x years interaction and $\mathbf{e}$ vector of residuals. $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{V}$, $\mathbf{T}$ and $\mathbf{R}$ is incidence matrix.

Quality analyzes were carried out with the parameters CR (Call Rate) and MAF (Minor Allele Frequency) equal to or greater than 90% and 5%, respectively, totaling 20,477 SNP markers.

2.2 Bayesian multi-trait genomic best linear unbiased prediction

The Bayesian multi-trait genomic best linear unbiased prediction (BMT-GBLUP) model used can be described as follows:

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

where, $\mathbf{y}$ is the vector of phenotypes (Y, VV, FS and NRN) ($t = 4$), X is the t x k incidence matrix of non-genetic effects; $\mathbf{b}$ is the $k \times 1$ vector of the non-genetic effects; $\mathbf{Z}$ is the $n \times m$ incidence matrix relating accessions with additive genomic effects; $\mathbf{g}$ is the $m \times 1$ vector of additive genomic effects, and e is the $t \times 1$ vector of residuals; and $\mathbf{e}$ is the $t \times 1$ vector of residuals. The $\mathbf{g}$ and $\mathbf{e}$ vectors were assumed to follow the independent multivariate Gaussian distributions $\mathbf{g} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{G})$ and $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_e \otimes \mathbf{I})$, respectively, where $\mathbf{G}$ is the genomic relationship matrix for genetic efects, $\mathbf{I}$ is the identity matrix for residuals, $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Sigma}_e$ are the ($t \times t$) variance-covariance matrices of genetic effects and residuals, respectively. Here, $\otimes$ indicates the Kronecker product. The $\mathbf{G}$ matrix was computed as $\mathbf{WW'} / \mathbf{2} \sum_{n=1}^{m} \boldsymbol{p_j}(\mathbf{1} - \boldsymbol{p_j})$, where $\mathbf{W}$ is an $n \times m$ matrix of centered SNP genotypes having values of $0 - 2p_j$ for zero copies of the reference allele, $1 - 2p_j$ for one copy of the reference allele, and $2 - 2p_j$ for two copies of the reference allele (VanRaden, 2008). Here, p_j corresponds to the allele frequency at SNP $j = 1, \ldots, m$. Flat priors were assigned to the intercepts and to the vector of fixed effects. Independent multivariate normal priors with null mean and inverse Wishart distributions, with hyperparameters ν and S, where ν is a scalar degrees of freedom and S is a positive-semi-defined symmetric matrix, for covariances matrices were assigned to the vectors of random additive genomic effects and residual effects.

Marginal posterior densities were obtained using a Markov Chain Monte Carlo (MCMC) approach with Gibbs sampling algorithm. Was used 1,200,000 MCMC samples with a burn-in of 50,000. The MCMC samples were thin interval equal to 50, resulting in 23,000 MCMC samples for inference. The posterior means of genetic values were used as inputs for inferring a trait network.

2.3 Bayesian networks

Bayesian networks describe conditional independence relationships between multivariate phenotypes (Korb & Nicholson, 2011). In this structure there are nodes, which would be the phenotypes, and the edges that connect the phenotypes if they are directly affected, the absence of an edge implies conditional independence between variables. The algorithm based on Hill Climbing (HC) scores was used, implemented in the R bnlearn package (Scutari, 2010) to infer the structure of the residual phenotypic Bayesian network for four (Y, VV, FS and NRN)

economic traits of coffee. Was computed the Bayesian information criterion (BIC) score afer each edge removal in the algorithm to infer their relative contribution to the overall BIC score of the network and estimated the strength and uncertainty of direction of each edge probabilistically by bootstrapping (n = 50,000 bootstraping samples). An edge strength ≥ 80% was used to select only high-confidence relationships.

2.4 Multi-trait MTM-GWAS

MTM-GWAS analyzes were performed using the SNP Snappy strategy (Meyer & Tier, 2012) implemented in the mixed model package WOMBAT (Meyer, 2007), according to the following model, which did not consider the inferred network structure:

$$\mathbf{y} = \mathbf{Ws} + \boldsymbol{X_b} + \boldsymbol{Z_g} + \mathbf{e}, \tag{3}$$

where $\mathbf{y}$ is the vector of phenotypes $(t = 5)$, $\mathbf{W}$ is the n x t by t matrix of genotype codes of SNP marker j, $\mathbf{s}$ is the $\mathrm{t} \times 1$ vector of direct effects for SNP marker j, and other terms were previously described. Variance-covariance structures were assumed the same as for Eq. (1). Was ftted MTM-GWAS for each SNP individually was fitted to obtain the following vector of marker estimates for each trait: $\mathbf{s} = [\mathbf{s_Y}, \mathbf{s_{VV}}, \mathbf{s_{FS}}, \mathbf{s_{NRN}}]$. A t statistic was used to obtain P-values: $\mathbf{T_{ij}} = \mathbf{s_j}/\mathbf{se}(\mathbf{s_j})$, where s is the point estimate of the jth SNP direct effect and $\mathrm{se}(s_j)$ is its standard error. The q-values were obtained by correcting the P-values for bonferroni protection with a significance level of 0.01.

2.5 Structural equations model – GWAS

The structured equation model manages to relate the network to the various phenotypes involving recursive effects. The use of SEM-GWAS was conducted using the SNP Snappy Strategy (Meyer & Tier, 2012) implemented in the mixed model package WOMBAT (Meyer, 2007). The SEM model described in Gianola and Sorensen was extended to GWAS according to Momen et al. (2018) and Momen et al. (2019).

$$\mathbf{y} = \boldsymbol{\Lambda}\mathbf{y} + \mathbf{Ws} + \mathbf{Zg} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{y}$ is the vector of phenotypes ($\mathrm{t} = 4$), and $\boldsymbol{\Lambda}$ is a $\mathrm{t} \times \mathrm{t}$ matrix of regression coefficients (structural coefficients) based on the learned structure from the Bayesian network using the residuals:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0 & \mathrm{I}_2\,\lambda_{\mathrm{VV}\to\mathrm{Y}} & \mathrm{I}_3\,\lambda_{\mathrm{FS}\to\mathrm{Y}} & \mathrm{I}_3\,\lambda_{\mathrm{NRN}\to\mathrm{Y}} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \mathrm{I}_2\,\lambda_{\mathrm{VV}\to\mathrm{NRN}} & 0 & 0 \end{bmatrix}$$

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

$$\psi = \begin{bmatrix} \sigma^2_{\mathrm{e(Y)}} & 0 & 0 & 0 \\ 0 & \sigma^2_{\mathrm{e(VV)}} & 0 & 0 \\ 0 & 0 & \sigma^2_{\mathrm{e(FS)}} & 0 \\ 0 & 0 & 0 & \sigma^2_{\mathrm{e(NNR)}} \end{bmatrix}.$$

All analyzes followed a routine that can be seen in figure 1.

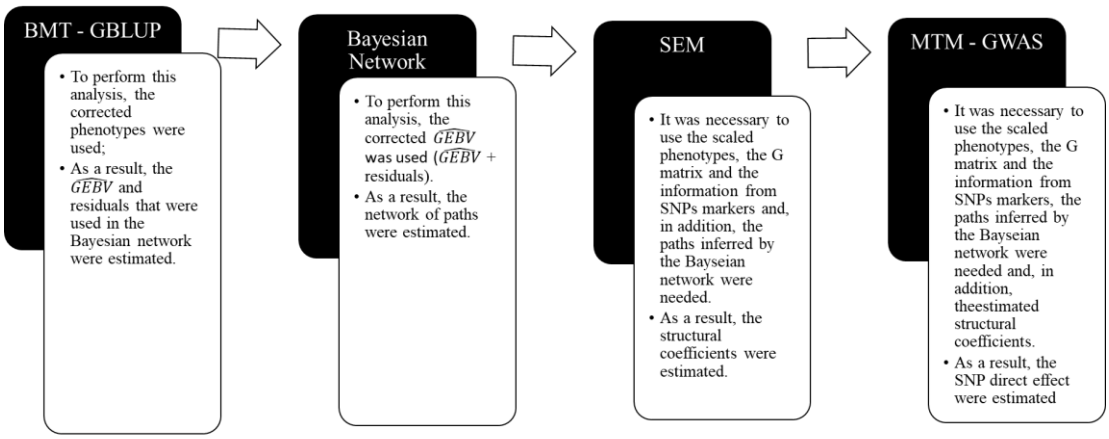


Figure 1: Flowchart detailing the analysis procedure.

The structural coefficients represented the size of the edge effect between phenotypes in the Bayesian network, so that the direct and indirect effects of the SNP effect could be compensated. While MTM-GWAS uses the effect of SNP as a direct effect, SEM considers it to be the direct effect of SNP, the indirect effects for the same SNP are obtained by those mediated by up-stream traits in the phenotypic network. The calculation of indirect effects

based on the multiplication of path coefficients for each path linking the SNP to an associated variable and then adding all these paths Mi et al. (2010) and Jiang et al. (2013). Thus, the general effect of the SNP is the sum of the direct and indirect effects sought for an analyzed characteristic.

The knowledge of direct and indirect effects is of great importance for the selection phase in breeding programs, whether plants or animals, which according to Valente et al. (2013) it is not possible using just MTM-GWAS. Thus, was used the results obtained with this methodology so that we could select markers that reflected significant effects on the characteristics under study.

3 RESULTS

3.1 Phenotypic correlations and Bayesian network structure

Descriptive statistics for the traits investigated are reported in Table 1. Average values were 5.19 liter/plant (4.76, 5.59) for Y, 7.35 (2.07, 7.47) for VV, 2.32 (1.99, 2.37) for FS, and 8.62 (7.19, 8.89) for NRN. Values in parentheses show lower and upper bounds of the highest 95% probability density regions (HPD95) obtained from the estimated marginal densities are given in parantheses

Genomic, residual correlations and heritability estimates obtained with a multi-trait Bayesian GBLUP model are reported in Table 1. No genomic correlation was obtained. Among residual correlations, we found relevant positive correlations between FS and Y (0.30) and between NRN and Y (0.38). Heritability estimates were moderate for VV (0.39) and FS (0.61), and low for Y (0.14) and NRN (0.13).

Table 1: Genomic (upper triangular) and residual (lower triangular) correlations, and genomic heritabilities (diagonals) for the coffee traits and their respective HPD (in parenthesis).

	Y	**VV**	**FS**	**NRN**
Y	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)

Y: yield; VV: vegetative vigor; FS: fruit size; NRN: number of reproductive nodes.
Relevant correlations (HPD95 not including 0) are highlighted in bold.

Bayesian network structure learning algorithms were applied to the residual plus breeding value vector of the Bayesian GBLUP analysis to identify dependencies between phenotypes. The results obtained with the HC algorithm are showed in the Figure 2.

Direction values represent the probability of the arc pointing to a particular node, and strength values represent

In this network, we found a direct dependence from VV to NRN (68% of bootstrap samples and 100% of strength), NRN to Y (100% of bootstrap samples and 81% of strength), VV to Y (100 % of bootstrap samples and 99% of strength) and FS to Y (100% of bootstrap samples and 82% of strength). The indirect path between VV and Y was mediated by the NRN.

The greatest decrease in BIC was observed when removing the VV → NRN arcs, suggesting that this path may play the most important role in the network (Table 2).

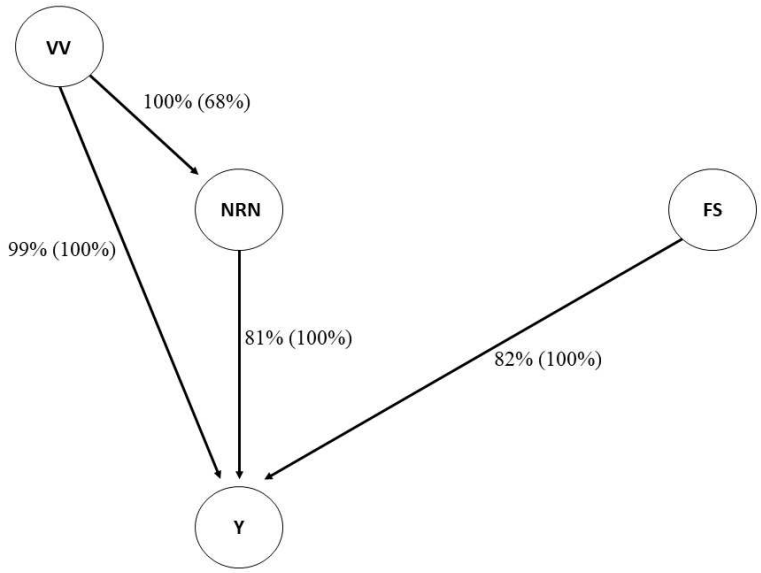


Figure 2: Network structure inferred from the vector of the residuals using the Hill-Climbing (HC) algorithm. Network structure inferred combining the results obtained with HC algorithm. Structure learning test was performed with 50,000 bootstrap samples. The percentages reported beside the edges indicate the proportion of the bootstrap samples supporting the edge and (in parentheses) the proportion having the direction shown.

Table 2: Bayesian Information Criterion (BIC) score for the Hill Climbing (HC) algorithm and path coefficients derived from the structural equation models.

BIC (a)	**Path**	**BIC (b)**	**Path coefficient (λ)**

-1395.29	VV → NRN	-15.19	0.0351
	VV → Y	-15.09	0.0047
	NRN → Y	-2.37	-0.0438
	FS → Y	-2.54	-0.0338

(a) Bayesian information criterion score (BIC) for 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 showed.

Y: yield; VV: vegetative vigor; FS: fruit size; NRN: number of reproductive nodes

3.2 Structural equation coefficients

Using the Bayesian network technique, it was possible to model the interrelationships between the four characteristics (Y, VV, NRN and FS), which enabled the construction of the DAG (Direct Acyclic Graphic), as can be seen in Figure 2. Using the SEM technique, it was possible to estimate the structural coefficients for each path, which enabled the estimation of the SNP effects. Table 2 shows the estimates of the structural coefficients. The coefficients of the NRN → Y and FS → Y paths were negative, while VV → NRN and VV → Y were positive. The coefficient referring to the NRN → Y path had the highest value, while VV → Y had the lowest coefficient.

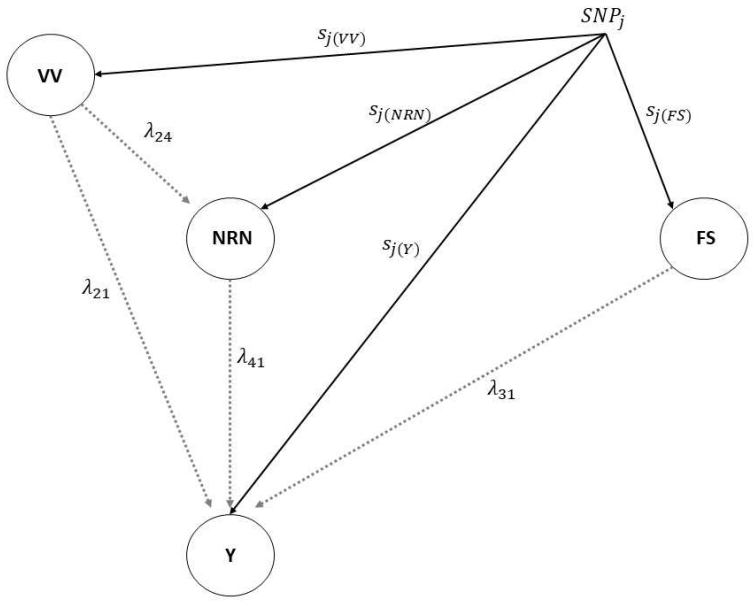


Figure 3: Figure for path analysis of SNP effects for four coffee-related traits. Y: yield; NRN: number of reproductive nodes; FS: fruit size; VV: vegetative vigor. The gray dashed arrows indicate the direction of relationship according to the learned causal structure. λ_{24}: VV → NRN;

λ_{21}: VV → Y; λ_{41}: NRN → Y; λ_{31}: FS → Y. The black arrows correspond to the direct effect of SNPj on the trait.

3.3 Partitioning of SNP effects

Using SEM-GWAS, it was possible to partition the effects of SNP into direct and one or more indirect effects. Manhattan plots for decomposition of the SNP effect are shown in Figs. 4-7.

3.3.1 Yield

λ_{24}: VV → NNR; λ_{21}: VV → Y; λ_{41}: NNR → Y; λ_{31}: FS → Y

Overall SNP effects for Y could be partitioned into one direct effect and four indirect effects (Fig. 1): (1) VV → Y, (2) NRN → Y, (3) FS → Y and (4) VV → NRN → Y. VV, NRN and FS influenced Y through an indirect path with structural coefficient λ_{21} (0.0047), λ_{41} (-0.0438) and λ_{31} (-0.0338). VV also indirectly contributed to NRN, which in turn affected Y, represented by the product of the coefficients λ_{24} x λ_{41} (0.0351 x -0.0438 = -0.0015). The contribution can be seen in Fig. 4.

$$Direct_{s_j \to y1_Y} = s_{j(y1_Y)}$$

$$Indirect(1)_{s_j \to y1_Y} = \lambda_{21} s_{j(y2_{VV})}$$

$$Indirect(2)_{s_j \to y1_Y} = \lambda_{41} s_{j(y4_{NRN})}$$

$$Indirect(3)_{s_j \to y1_Y} = \lambda_{31} s_{j(y3_{FS})}$$

$$Indirect(4)_{s_j \to y1_Y} = \lambda_{21}\lambda_{41} s_{j(y2_{VV})}$$

$$\begin{aligned} Total_{s_j \to y1_Y} &= Direct_{s_j \to y1_Y} + Indirect(1)_{s_j \to y1_Y} + Indirect(2)_{s_j \to y1_Y} \\ &+ Indirect(3)_{s_j \to y1_Y} + Indirect(4)_{s_j \to y1_Y} \\ &= s_{j(y1_Y)} + \lambda_{21} s_{j(y2_{VV})} + \lambda_{41} s_{j(y4_{NRN})} + \lambda_{31} s_{j(y3_{FS})} + \lambda_{24}\lambda_{41} s_{j(y2_{VV})} \end{aligned}$$

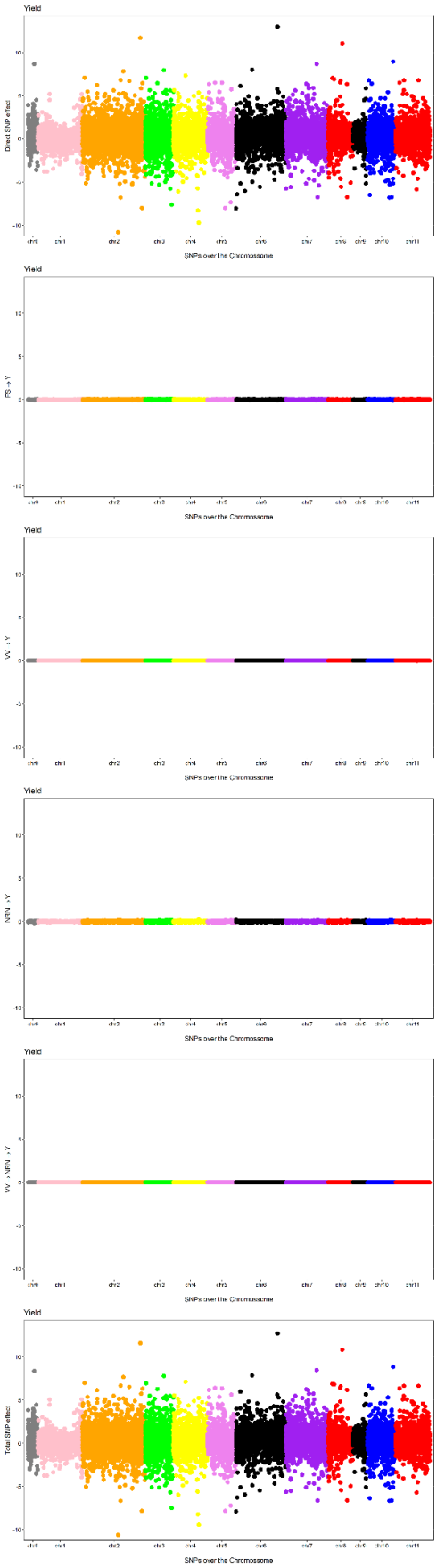

Yield
Direct SNP effect
SNPs over the Chromosome
Yield
FS x Y
SNPs over the Chromosome
Yield
VV x Y
SNPs over the Chromosome
Yield
NRN x Y
SNPs over the Chromosome
Yield
VV x NRN x Y
SNPs over the Chromosome
Yield
Total SNP effect
SNPs over the Chromosome

Figure 4: Manhattan plots for SNP effects on yield obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. VV: vegetative vigor; NRN: number of reproductive nodes; Y: yield

3.3.2 Vegetative Vigor

In the case of VV, the Bayesian network algorithm did not identify any mediator trait (Fig. 1). Therefore, the genomic architecture of VV was seemingly controlled only by direct SNP effects, i.e., the total effect of the jth SNP on VV corresponds to its own direct effect (Fig. 5).

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

$$Total_{s_j \to y2_{VV}} = Direct_{s_{j(y2_{VV})}} = s_{j(y2_{VV})}$$

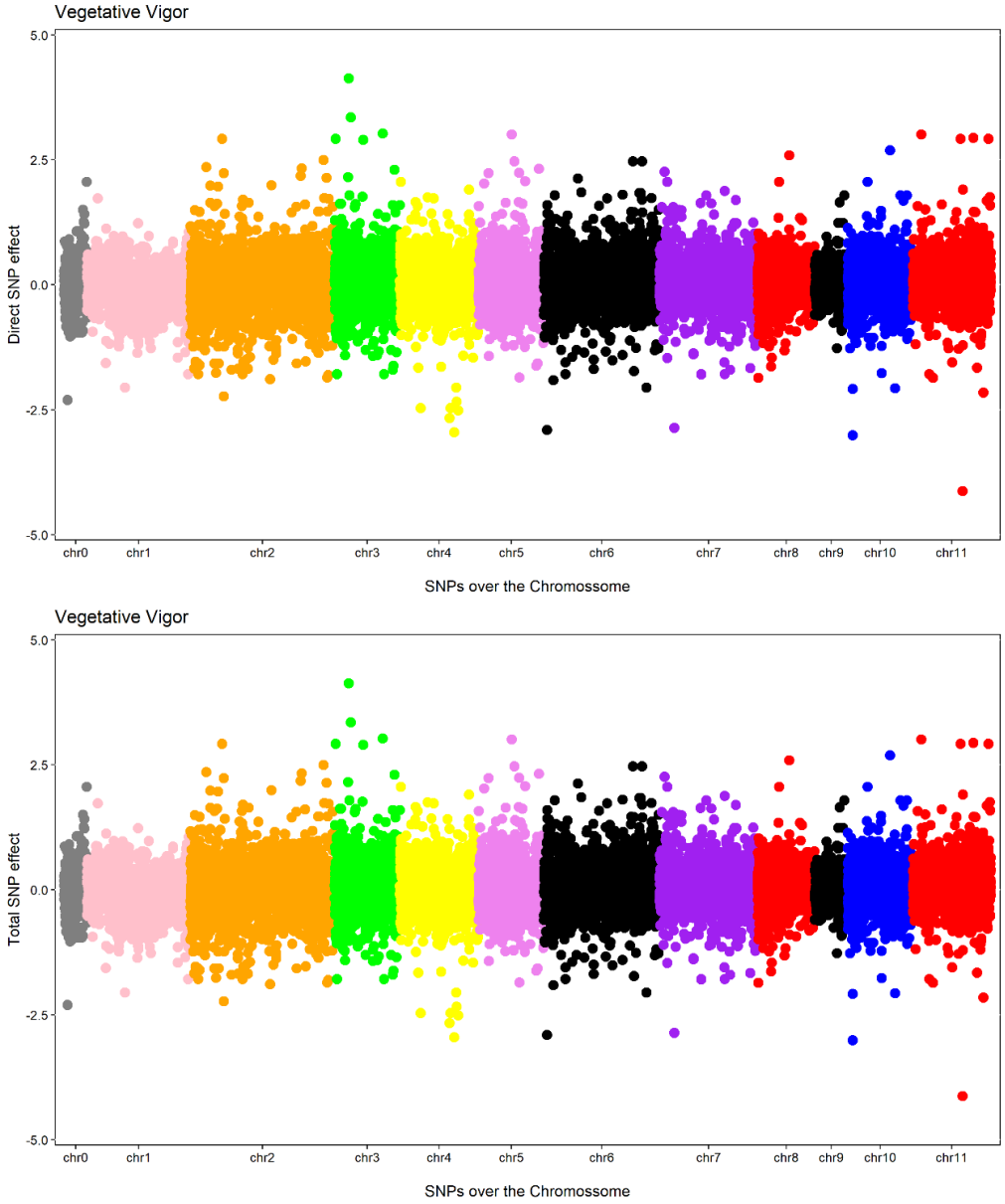


Figure 5: Manhattan plots for SNP effects on number of reproductive nodes obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. VV: vegetative vigor.

3.3.3 Fruit Size

In the case of FS, the Bayesian network algorithm did not identify mediator trait (Fig. 1). Therefore, the genomic architecture of FS was seemingly controlled only by direct SNP effects, i.e., the total effect of the jth SNP on FS corresponds to its own direct effect (Fig. 6).

$$Direct_{s_j \to y3_{FS}} = s_{j(y3_{FS})}$$

$$Total_{s_j \to y3_{FS}} = Direct_{s_{j(y3_{FS})}} = s_{j(y3_{FS})}$$

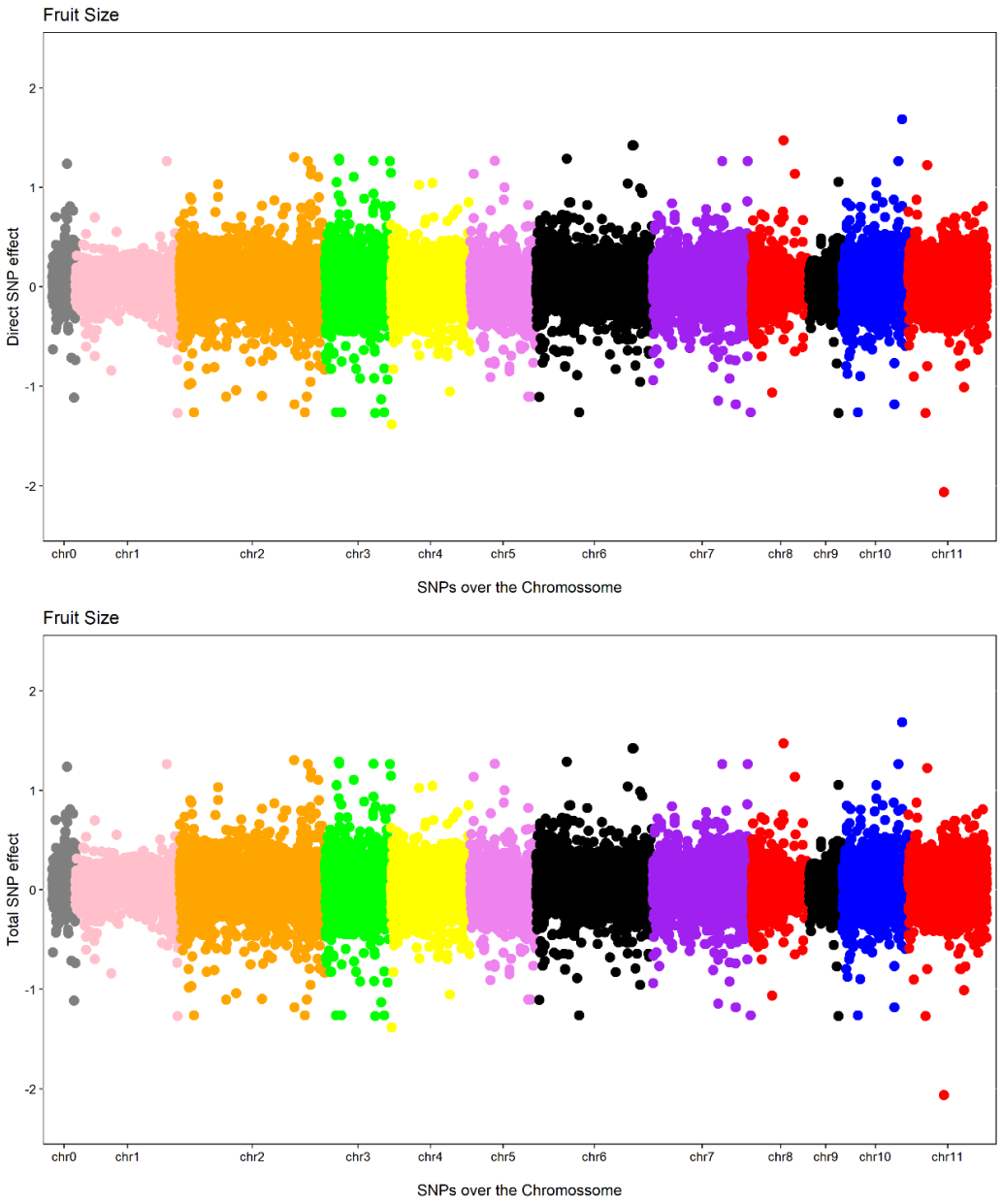


Figure 6: Manhattan plots for SNP effects on fruit size obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. FS: fruit size.

3.3.4 Number of Rproductive Nodes

The overall SNP effect on NRN was decomposed into one direct effect and one indirect effect mediated by VV (VV → NRN) with a structural coefficient λ_{24} (0.0351). The contribution to SNP effects on NRN mediated by VV (Fig. 7).

$$Direct_{s_j \to y4_{NRN}} = s_{j(y4_{NRN})}$$

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

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

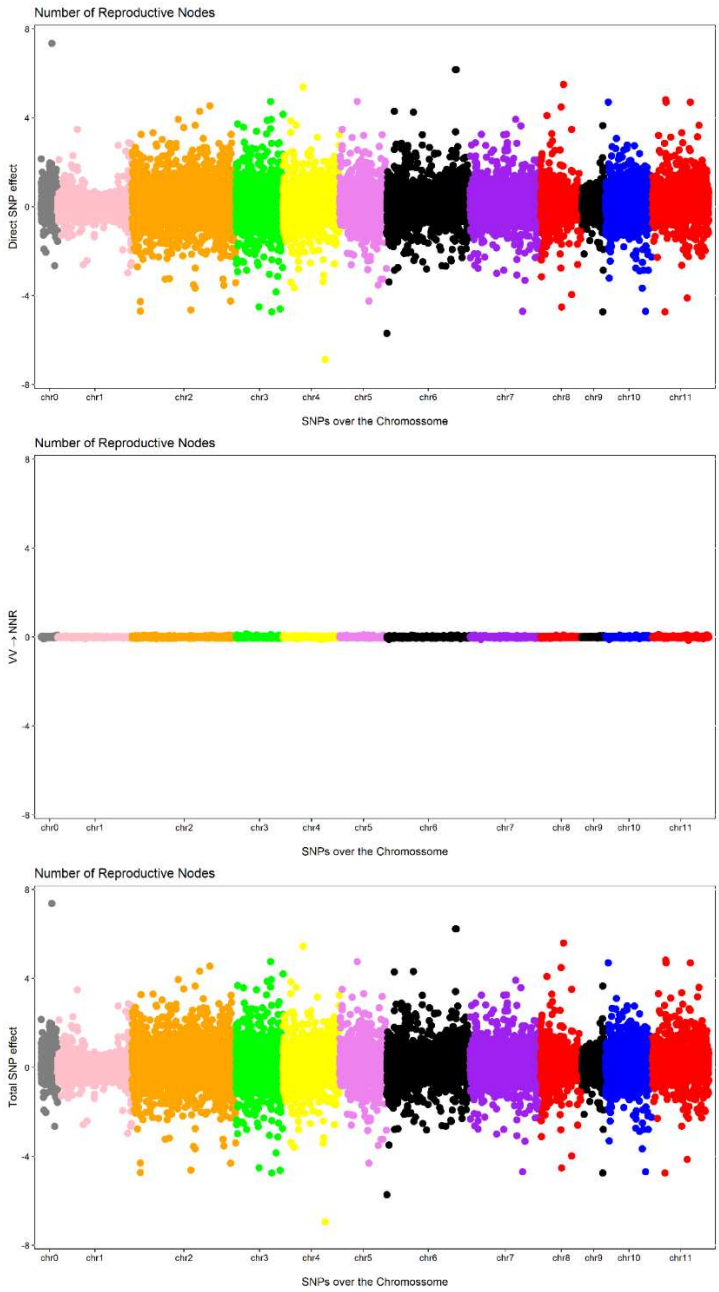


Figure 7: Manhattan plots for SNP effects on number of reproductive nodes obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. VV: vegetative vigor; NRN: number of reproductive nodes.

Was compared the direct and indirect SNP effects with the total SNP effects for NRN and Y. Direct SNP effects were positively highly correlated (R^2 > 0.98) with total SNP effects for all traits. For the indirect SNP effects with total SNP effects were positively correlated for VV → NNR (0.02) and VV → Y (0.03), and negatively correlated for NRN → Y (0.72), FS → Y (0.14) and VV → NNR → Y (0.03), as seen in the attachments.

3.4 Genome-Wide Association Study for Yield, Vegetative Vigor, Fruit Size and Number of Reproductive Nodes

Two hundred and ninety-seven SNP were statistically significant, however, seven are allocated in the Unchr (Uncharacterizad chromosome), the Unchr is constituted by a set of scaffolds with disordered sequences, according to information found in this region not discussed in this work. Thus, 290 significant SNP were obtained, where five are related to the NRN characteristic and two hundred and eighty-five to the Y characteristic ($q < 0.01$) (Table 3). These SNP are distributed on chromosomes 1 (Chr 1) to 11 (Chr 11) (Figure 4 to 7).

Table 3: SNP with significant associations ($q<0.01$), chromosome and position associated with Y, VV, FS and NRN.

SNP	Chr	Position	q-value
V5938	Unchr	177	4.80E-03
V3683	Unchr	4424	7.36E-04
V1324	Unchr	14706	1.07E-03
V2880	Unchr	24645	2.61E-03
V887	Unchr	402268	2.70E-03
V888	Unchr	402295	4.07E-05
V1309	Unchr	453050	9.09E-03
V2681	1	3925767	1.87E-03
V2102	1	8299841	7.02E-04
V773	1	8922851	4.05E-04
V1101	1	9651201	3.79E-03
V2830	1	10231888	3.76E-04
V510	1	11268905	5.82E-03
V2856	1	12667165	9.96E-05
V2857	1	12667184	3.29E-05
V2861	1	12874767	1.10E-03
V2935	1	14094952	9.59E-03
V2032	1	33500866	2.88E-04
V3116	1	34314786	9.53E-03
V3117	1	34382600	9.69E-03
V3244	1	36669403	2.47E-04
V1656	1	37086050	2.95E-05
V1389	1	37493921	3.50E-04
V1393	1	37493976	7.27E-04
V3534	1	38197381	4.86E-03
V3598	1	38972375	4.04E-03
V3635	1	39279256	9.00E-03
V3638	1	39283941	4.10E-04

V3664	1	39458195	3.79E-03
V3670	1	39517087	2.32E-06
V3671	1	39517161	7.83E-05
V3674	1	39574545	3.80E-05
V3379	1	40487725	2.48E-03
V3469	1	40800016	8.12E-03
V3904	1	41046162	8.79E-03
V3905	1	41046176	4.11E-03
V3488	1	41151580	5.20E-03
V3489	1	41151609	4.09E-03
V3526	1	41438835	7.42E-03
V3537	1	41565994	3.65E-04
V3563	1	41875031	5.61E-04
V3571	1	41913639	2.90E-05
V3578	1	42059867	2.72E-03
V3579	1	42059892	1.84E-04
V3585	1	42062620	3.01E-04
V3589	1	42094959	2.10E-05
V3605	1	42356201	1.85E-04
V3618	1	42412779	5.54E-06
V3621	1	42448177	2.23E-05
V3631	1	42522006	4.25E-05
V3632	1	42522020	6.02E-04
V3639	1	42591761	2.73E-04
V3652	1	42649531	5.12E-04
V3657	1	42695867	3.84E-03
V3659	1	42704583	4.23E-04
V3660	1	42709732	2.73E-04
V3661	1	42709765	2.13E-04
V4283	1	44680136	7.78E-03
V4322	1	44854051	2.25E-03
V4493	1	46145891	2.44E-03
V4103	1	46415138	3.55E-04
V4117	1	46448319	1.10E-03
V1163	1	46919213	1.28E-03
V4807	1	48586572	8.22E-04
V4605	1	48935415	4.10E-03
V1306	1	48980848	3.19E-03
V4674	1	49440979	9.29E-03
V4736	1	49992250	3.42E-04
V4737	1	49992264	2.15E-03
V4808	1	50540487	3.67E-05
V7568	2	310332	2.06E-05
V7591	2	522419	7.94E-04
V7682	2	1546167	1.02E-03
V7711	2	1951604	8.23E-04

V7733	2	2228026	3.51E-03
V7771	2	2667752	6.19E-03
V7772	2	2667766	5.41E-04
V7809	2	3390127	6.02E-03
V7848	2	4210645	7.96E-04
V8141	2	8227345	5.85E-03
V8142	2	8227346	4.49E-03
V8220	2	9179904	4.98E-03
V8271	2	9781605	1.41E-03
V8277	2	9803905	2.54E-03
V8278	2	9805963	8.43E-03
V8289	2	10266215	2.36E-03
V8294	2	10304755	6.84E-04
V8322	2	10520999	2.88E-04
V8328	2	10592278	6.44E-05
V8336	2	10842468	1.17E-03
V8349	2	11200423	8.46E-03
V8368	2	11368300	4.91E-03
V8370	2	11479801	3.70E-04
V8387	2	12054716	3.97E-03
V8389	2	12063855	3.90E-05
V8391	2	12120840	8.84E-04
V8419	2	12256903	3.83E-03
V8429	2	12318079	1.04E-05
V8430	2	12318082	8.76E-04
V8398	2	12412575	2.45E-04
V8405	2	12557508	7.98E-03
V8433	2	12943209	1.46E-04
V8489	2	14144922	5.97E-03
V8555	2	14503278	7.37E-03
V8514	2	14863253	5.62E-03
V8592	2	15122048	4.54E-03
V8593	2	15122115	7.68E-03
V8608	2	15199584	4.41E-07
V8609	2	15199593	8.79E-05
V8610	2	15202975	1.86E-05
V8611	2	15202976	9.93E-04
V8625	2	15347458	7.38E-04
V8631	2	15384134	6.10E-03
V8673	2	16168941	3.27E-03
V8676	2	16168977	4.40E-03
V8652	2	16214191	8.44E-03
V8664	2	16421319	1.13E-03
V8665	2	16421369	2.00E-03
V8668	2	16499208	2.11E-03
V8669	2	16499212	8.21E-03

V8677	2	16560250	2.71E-04
V8680	2	16571219	1.22E-03
V8687	2	16594128	1.67E-03
V8743	2	16907517	4.17E-04
V8744	2	16907560	2.86E-03
V8587	2	17847283	6.22E-03
V8977	2	20172274	9.11E-03
V8849	2	20202098	5.35E-03
V8868	2	20395684	4.45E-04
V8895	2	20806412	8.81E-03
V8896	2	20810259	2.36E-03
V8905	2	21060957	4.81E-04
V9011	2	21195891	3.55E-03
V995	2	27137519	8.60E-03
V354	2	27616187	3.45E-03
V9189	2	31083032	4.69E-03
V9295	2	33094940	2.62E-04
V9247	2	35730459	6.43E-03
V9252	2	35730541	7.83E-03
V1359	2	36791605	8.96E-04
V9387	2	51481880	1.06E-03
V147	2	53474348	4.14E-05
V1716	2	53584949	1.99E-03
V9551	2	56740972	7.14E-03
V1259	2	56933743	8.93E-04
V9634	2	57478470	3.77E-03
V1979	2	57757361	4.77E-05
V9697	2	58842880	7.61E-03
V9684	2	59202346	7.23E-04
V9666	2	59551188	2.48E-04
V9652	2	59888512	3.78E-03
V9802	2	60101536	9.05E-04
V9964	2	61762987	3.85E-03
V9932	2	62076582	9.04E-04
V10014	2	62786557	5.97E-04
V10155	2	64080402	1.44E-03
V10177	2	64504213	8.29E-04
V10178	2	64504222	5.61E-04
V9708	2	64561829	2.84E-03
V10011	2	68177619	9.11E-03
V10230	2	70274536	5.13E-03
V10231	2	70279929	9.65E-04
V10233	2	70279995	4.76E-03
V1360	3	3754775	9.75E-03
V958	3	7262349	2.30E-05
V959	3	7262380	1.90E-04

V961	3	7262421	1.97E-04
V1790	3	9517428	3.81E-03
V491	3	11695179	9.25E-04
V2036	3	15592880	6.36E-05
V2037	3	15592885	4.42E-05
V1507	3	22500714	2.64E-03
V1799	3	22511327	2.22E-04
V962	3	25515454	1.89E-03
V2928	3	28086198	1.59E-04
V170	3	28660032	5.92E-04
V82	3	34556827	5.12E-05
V83	3	34556844	1.88E-04
V2039	4	5335318	3.58E-03
V1729	4	10891858	5.51E-03
V1527	4	12805764	1.60E-03
V86	4	17395815	3.81E-03
V927	4	36103351	6.39E-05
V6691	4	38996008	3.76E-05
V1448	5	11165107	8.14E-03
V213	5	18861837	4.21E-05
V1395	5	24764563	1.83E-03
V1314	5	25578972	4.47E-05
V1315	5	25578998	5.61E-04
V2357	5	36230080	3.68E-03
V1405	5	37804264	3.58E-03
V1981	6	18199439	6.72E-03
V1983	6	18199494	5.59E-05
V6318	6	26364086	2.44E-03
V6319	6	26364122	8.98E-04
V1094	6	31352169	5.57E-03
V685	6	33780141	5.63E-04
V1505	6	34373031	3.34E-03
V7	6	38649604	9.60E-04
V1912	6	42343792	5.05E-04
V703	6	43593076	1.38E-03
V1340	6	44683598	6.38E-03
V10040	6	54040779	5.46E-03
V10038	6	54076570	2.56E-03
V2356	7	1229956	7.89E-03
V856	7	2564680	9.77E-03
V941	7	6028346	3.01E-03
V862	7	8007514	1.20E-03
V890	7	16431338	3.52E-03
V75	7	33706981	8.33E-05
V98	8	5120345	2.24E-03
V99	8	5120353	1.29E-03

V406	8	5542367	9.14E-03
V345	8	22264345	1.76E-04
V2028	9	4319534	5.42E-03
V9147	9	5169310	4.65E-04
V1583	9	5251763	8.20E-04
V1012	9	11619464	8.42E-04
V1411	9	26900826	8.03E-03
V1470	9	33200105	8.33E-03
V4918	10	2553024	6.76E-03
V4919	10	2553056	6.36E-03
V4920	10	2553077	3.15E-04
V4921	10	2594407	2.96E-04
V4922	10	2594411	8.57E-04
V5156	10	6590142	2.08E-03
V5184	10	7442209	6.16E-03
V5963	10	7460043	3.68E-04
V1944	10	10066121	3.84E-04
V5450	10	11872547	3.87E-03
V214	10	24056448	2.22E-04
V215	10	24056497	4.77E-04
V216	10	24056507	2.86E-05
V217	10	24056512	8.35E-03
V709	10	27791481	1.33E-03
V712	10	27791560	9.33E-03
V2931	10	30642093	3.90E-03
V5514	10	34808941	3.77E-04
V5361	10	36968487	2.90E-04
V5336	10	37712589	8.61E-03
V220	10	38001494	5.64E-03
V5699	10	40827076	3.88E-03
V5700	10	40827089	2.14E-03
V5564	10	42635791	3.17E-03
V1286	10	42738120	9.67E-03
V5809	10	44037441	4.26E-03
V5810	10	44037450	4.26E-03
V5811	10	44037473	3.02E-04
V5815	10	44150033	8.87E-05
V5816	10	44150035	5.16E-05
V5820	10	44183992	3.17E-04
V5824	10	44184019	6.84E-03
V5838	10	44457920	5.67E-03
V5877	10	44741658	1.98E-03
V6288	11	1860064	2.60E-05
V6289	11	1905439	1.14E-05
V6291	11	1905485	1.25E-03
V6306	11	2236120	2.55E-03

V6307	11	2236125	7.62E-04
V6308	11	2236136	2.37E-05
V6309	11	2236140	7.54E-03
V6038	11	4709133	2.50E-03
V1811	11	6804560	1.22E-03
V1812	11	6804638	7.05E-03
V772	11	7065738	5.12E-04
V6304	11	7623511	4.07E-04
V6305	11	7623569	1.50E-03
V6264	11	12674994	3.39E-04
V872	11	13432126	6.83E-04
V6215	11	22256731	7.74E-05
V6226	11	22385262	1.66E-03
V6477	11	23700363	1.94E-03
V6790	11	26648349	7.52E-03
V6791	11	26648392	7.04E-03
V6820	11	27900161	3.06E-03
V6880	11	28534901	1.52E-03
V6969	11	29381221	9.55E-05
V7006	11	29894818	6.13E-04
V7053	11	30452457	2.81E-03
V1074	11	30493159	5.59E-03
V7109	11	31449939	5.08E-04
V6549	11	31684338	7.57E-03
V6594	11	32213390	1.03E-05
V7191	11	32300796	2.46E-03
V6642	11	32681014	2.80E-03
V6690	11	33844511	1.36E-03
V6857	11	36152396	2.04E-03
V6886	11	36615617	6.55E-04
V6899	11	36768547	3.40E-05
V6900	11	36768563	3.14E-04
V6903	11	36768655	3.04E-05
V6927	11	36901542	8.16E-04
V7409	11	40608337	9.46E-03
V7438	11	41084720	2.43E-03
V7443	11	41135712	5.62E-06
V7444	11	41135773	1.42E-05
V7360	11	42289762	5.14E-03

C. arabica is an allotetraploid from *C. canephora* and *C. eugenioides* (Lashermes et al., 1999), so its genome is divided into two subgenomes, so the front of each SNP marker code is preceded by "c" and "e", referring to *C. canephora* and *C. eugenioides*, respectively, as can be seen in the table 4.

Table 4: Functional annotation of SNP insert in genes for Number of Reproductive Nodes and Yield.

SNP	Chr	Position	Gene	Functional annotation	Trait
V2681_c	1	3925767	LOC113700785	Uncharacterized	Y
V773_c	1	8922851	LOC113712526	Cytochrome P450 81E8-like	Y
V2830_c	1	10231888	LOC113737716	Uncharacterized	Y
V510_e	1	11268905	LOC113700991	IRK-interacting protein-like	Y
V2935_e	1	14094952	LOC113701199	Serine/threonine-protein kinase Nek6-like	Y
V3116_c	1	34314786	LOC113724086	protein GrpE-like	NRN
V3116_e	1	34314786	LOC113703827	G2/mitotic-specific cyclin S13-7-like	NRN
V3117_c	1	34382600	LOC113739999	phytochromobilin:ferredoxin oxidoreductase, chloroplastic-like	NRN
V3117_e	1	34382600	LOC113703885	alcohol dehydrogenase-like 7	NRN
V1656_e	1	37086050	LOC113689416	aluminum-activated malate transporter 2-like	Y
V1389_e	1	37493921	LOC113705580	leucine-rich repeat-containing protein ODA7-like	Y
V1393_e	1	37493976	LOC113706063	leucine-rich repeat-containing protein ODA7-like	Y
V3534_e	1	38197381	LOC113706812	poly [ADP-ribose] polymerase 3-like	Y
V3598_c	1	38972375	LOC113741358	spermidine hydroxycinnamoyl transferase-like	Y
V3598_e	1	38972375	LOC113707094	SRSF protein kinase 2-like	Y
V3635_e	1	39279256	LOC113707094	shugoshin-1-like	Y
V3638_c	1	39283941	LOC113725576	indole-3-acetaldehyde oxidase-like	Y
V3638_e	1	39283941	LOC113695297	uncharacterized	Y
V3664_c	1	39458195	LOC113741525	receptor-like protein Cf-9 homolog	Y
V3664_e	1	39458195	LOC113707272	ABC transporter F family member 1-like	Y
V3670_e	1	39517087	LOC113707286	ABC transporter G family member 3	Y
V3671_e	1	39517161	LOC113707286	ABC transporter G family member 3	Y
V3674_e	1	39574545	LOC113707345	SNF1-related protein kinase regulatory subunit beta-3	Y
V3379_e	1	40487725	LOC113708346	pentatricopeptide repeat-containing protein	Y

				At2g27800, mitochondrial-like	
V3469_c	1	40800016	LOC113726210	ubiquitin-conjugating enzyme E2 variant 1C-like	Y
V3469_e	1	40800016	LOC113708681	probable protein phosphatase 2C 4	Y
V3904_e	1	41046162	LOC113708932	uncharacterized	Y
V3905_e	1	41046176	LOC113708932	uncharacterized	Y
V3488_e	1	41151580	LOC113709139	beta-Amyrin Synthase 2-like	Y
V3489_e	1	41151609	LOC113709139	beta-Amyrin Synthase 2-like	Y
V3526_c	1	41438835	LOC113725935	uncharacterized	Y
V3537_c	1	41565994	LOC113726673	poly [ADP-ribose] polymerase 3-like	Y
V3537_e	1	41565994	LOC113709643	uncharacterized	Y
V3571_c	1	41913639	LOC113726852	uncharacterized	Y
V3578_c	1	42059867	LOC113726927	uncharacterized	Y
V3579_c	1	42059892	LOC113726927	uncharacterized	Y
V3585_c	1	42062620	LOC113726938	hyoscyamine 6-dioxygenase-like	Y
V3585_e	1	42062620	LOC113710247	eukaryotic translation initiation factor 3 subunit G-like	Y
V3605_c	1	42356201	LOC113727118	vacuolar protein sorting-associated protein 41 homolog	Y
V3605_e	1	42356201	LOC113710693	protein NRT1/ PTR FAMILY 8.2-like	Y
V3621_c	1	42448177	LOC113727154	cyclin-dependent kinases regulatory subunit 1	Y
V3631_c	1	42522006	LOC113727246	probable aquaporin TIP1-1	Y
V3632_c	1	42522020	LOC113727246	probable aquaporin TIP1-1	Y
V3639_e	1	42591761	LOC113711001	villin-1-like	Y
V3652_e	1	42649531	LOC113711078	vacuolar cation/proton exchanger 3-like	Y
V3657_c	1	42695867	LOC113727337	4-coumarate--CoA ligase-like 7	Y
V3659_c	1	42704583	LOC113727343	uncharacterized	Y
V3659_e	1	42704583	LOC113711170	protein O-glucosyltransferase 1-like	Y
V4283_e	1	44680136	LOC113713753	uncharacterized	Y
V4322_e	1	44854051	LOC113714036	basic blue protein-like	Y
V4493_c	1	46145891	LOC113730695	ERBB-3 BINDING PROTEIN 1-like	Y
V4103_c	1	46415138	LOC113729054	LOB domain-containing protein 12	Y
V4103_e	1	46415138	LOC113716136	homocysteine S-methyltransferase 2-like	Y

V4117_c	1	46448319	LOC113742966	malonyl CoA-acyl carrier protein transacylase	Y
V4807_e	1	48586572	LOC113719062	ABC transporter B family member 11-like	Y
V4605_c	1	48935415	LOC113733971	BSD domain-containing protein 1-like	Y
V1306_c	1	48980848	LOC113734013	methyl-CpG-binding domain-containing protein 4-like	Y
V4674_c	1	49440979	LOC113734653	cullin-1-like	Y
V4736_c	1	49992250	LOC113735447	protein DEHYDRATION-INDUCED 19 homolog 7-like	Y
V4737_c	1	49992264	LOC113735447	protein DEHYDRATION-INDUCED 19 homolog 7-like	Y
V4808_c	1	50540487	LOC113736006	dnaJ homolog subfamily B member 6-like	Y
V7568_e	2	310332	LOC113733380	protein EXPORTIN 1A-like	Y
V7591_c	2	522419	LOC113724807	UNC93-like protein 1	Y
V7682_c	2	1546167	LOC113724913	probable WRKY transcription factor 41	Y
V7682_e	2	1546167	LOC113729964	plastid division protein PDV1-like	Y
V7733_c	2	2228026	LOC113725012	uncharacterized	Y
V7733_e	2	2228026	LOC113730048	kinesin-like protein KIN-10C	Y
V7771_c	2	2667752	LOC113725069	probable methyltransferase PMT15	Y
V7772_c	2	2667766	LOC113725069	probable methyltransferase PMT15	Y
V7809_c	2	3390127	LOC113725152	BAG family molecular chaperone regulator 7-like	Y
V7848_c	2	4210645	LOC113725255	methylesterase 17	Y
V7848_e	2	4210645	LOC113730254	macro domain-containing protein VPA0103-like	Y
V8141_e	2	8227345	LOC113730710	receptor-like protein kinase FERONIA	Y
V8142_e	2	8227346	LOC113730710	receptor-like protein kinase FERONIA	Y
V8220_e	2	9179904	LOC113730801	double-stranded RNA-binding protein 2-like	Y
V8271_c	2	9781605	LOC113725874	transcription factor DYT1-like	Y
V8271_e	2	9781605	LOC113728538	uncharacterized	Y
V8277_e	2	9803905	LOC113730866	G-type lectin S-receptor-like serine/threonine-protein kinase At4g27290	Y

V8278_e	2	9803905	LOC113730866	G-type lectin S-receptor-like serine/threonine-protein kinase At4g27290	Y
V8277_c	2	9803905	LOC113725875	uncharacterized	Y
V8278_c	2	9805963	LOC113725876	DAG protein, chloroplastic-like	Y
V8289_c	2	10266215	LOC113725911	lysine-specific demethylase JMJ25-like	Y
V8289_e	2	10266215	LOC113730904	myosin-2-like	Y
V8294_c	2	10304755	LOC113725914	transcription factor MYB102-like	Y
V8322_c	2	10520999	LOC113725934	quinone oxidoreductase PIG3-like	Y
V8328_c	2	10592278	LOC113725942	ATP-dependent RNA helicase A-like	Y
V8328_e	2	10592278	LOC113730930	helicase-like transcription factor CHR28	Y
V8349_c	2	11200423	LOC113723900	pentatricopeptide repeat-containing protein At3g29230-like	Y
V8349_e	2	11200423	LOC113730977	uncharacterized	Y
V8368_c	2	11368300	LOC113726018	pentatricopeptide repeat-containing protein At2g33760-like	Y
V8370_c	2	11479801	LOC113723906	short-chain dehydrogenase reductase 3b-like	Y
V8387_c	2	12054716	LOC113726081	uncharacterized	Y
V8389_c	2	12063855	LOC113726082	glutaredoxin-C9-like	Y
V8419_e	2	12256903	LOC113731050	protein CHUP1, chloroplastic	Y
V8419_c	2	12256903	LOC113726092	uncharacterized	Y
V8429_e	2	12318079	LOC113731056	uncharacterized	Y
V8430_e	2	12318082	LOC113731056	uncharacterized	Y
V8398_c	2	12412575	LOC113726102	negative regulator of systemic acquired resistance SNI1-like	Y
V8398_e	2	12412575	LOC113728599	receptor kinase-like protein Xa21	Y
V8405_e	2	12557508	LOC113731072	uncharacterized	Y
V8433_c	2	12943209	LOC113724380	probable LRR receptor-like serine/threonine-protein kinase At3g47570	Y
V8433_e	2	12943209	LOC113731096	receptor kinase-like protein Xa21	Y
V8489_c	2	14144922	LOC113726226	probable LRR receptor-like serine/threonine-protein kinase RFK1	Y
V8555_e	2	14503278	LOC113731236	purple acid phosphatase 2	Y

V8514_c	2	14863253	LOC113726260	peptidyl-prolyl cis-trans isomerase CYP21-1-like	Y
V8514_e	2	14863253	LOC113731283	probable LRR receptor-like serine/threonine-protein kinase At2g16250	Y
V8592_c	2	15122048	LOC113726287	phosphatidylinositol/phosphatidylcholine transfer protein SFH3-like	Y
V8593_c	2	15122115	LOC113726287	phosphatidylinositol/phosphatidylcholine transfer protein SFH3-like	Y
V8592_e	2	15122048	LOC113729561	adenylate isopentenyltransferase 5, chloroplastic-like	Y
V8593_e	2	15122115	LOC113729561	adenylate isopentenyltransferase 5, chloroplastic-like	Y
V8608_c	2	15199584	LOC113726299	probable pinoresinol-lariciresinol reductase 3	Y
V8609_c	2	15199593	LOC113726299	probable pinoresinol-lariciresinol reductase 3	Y
V8608_e	2	15199584	LOC113731333	EEF1A lysine methyltransferase 4-like	Y
V8609_e	2	15199593	LOC113731333	EEF1A lysine methyltransferase 4-like	Y
V8610_e	2	15202975	LOC113731335	probable E3 ubiquitin-protein ligase ARI2	Y
V8611_e	2	15202976	LOC113731335	probable E3 ubiquitin-protein ligase ARI2	Y
V8625_e	2	15347458	LOC113731354	serine/threonine-protein kinase SMG1-like	Y
V8631_e	2	15384134	LOC113731363	diphthamide biosynthesis protein 3-like	Y
V8631_e	2	15384134	LOC113731362	diphthamide biosynthesis protein 3-like	Y
V8673_c	2	16168941	LOC113726423	autophagy-related protein 2-like	Y
V8676_c	2	16168977	LOC113726423	autophagy-related protein 2-like	Y
V8673_e	2	16168941	LOC113731461	stress enhanced protein 1, chloroplastic-like	Y
V8676_e	2	16168977	LOC113731461	stress enhanced protein 1, chloroplastic-like	Y
V8652_c	2	16214191	LOC113726426	protein CYPRO4-like	Y
V8652_e	2	16214191	LOC113731468	cell division cycle protein 123 homolog	Y
V8668_c	2	16499208	LOC113726461	DNA-directed RNA polymerases II, IV and V subunit 3-like	Y

V8669_c	2	16499212	LOC113726461	DNA-directed RNA polymerases II, IV and V subunit 3-like	Y
V8677_c	2	16560250	LOC113726472	probable LRR receptor-like serine/threonine-protein kinase At4g37250	Y
V8680_c	2	16571219	LOC113726474	cell division cycle protein 123 homolog	Y
V8680_e	2	16571219	LOC113731519	ATP-dependent Clp protease proteolytic subunit-related protein 1, chloroplastic	Y
V8687_c	2	16594128	LOC113726478	cinnamyl alcohol dehydrogenase 1-like	Y
V8743_e	2	16907517	LOC113731558	aspartate--tRNA ligase, chloroplastic/mitochondrial	NRN
V8743_c	2	16907517	LOC113726524	BTB/POZ domain-containing protein NPY2-like	NRN
V8744_e	2	16907560	LOC113731558	aspartate--tRNA ligase, chloroplastic/mitochondrial	NRN
V8744_c	2	16907560	LOC113726524	BTB/POZ domain-containing protein NPY2-like	NRN
V8587_e	2	17847283	LOC113731653	phosphatidylinositol 4-kinase alpha 1-like	Y
V8587_c	2	17847283	LOC113726636	uncharacterized	Y
V8977_e	2	20172274	LOC113731882	ATP-dependent DNA helicase Q-like 4A	Y
V8977_c	2	20172274	LOC113722748	uncharacterized	Y
V8849_e	2	20202098	LOC113731885	MLO-like protein 4	Y
V8868_c	2	20395684	LOC113725085	uncharacterized	Y
V8868_e	2	20395684	LOC113731911	uncharacterized	Y
V8895_e	2	20806412	LOC113728771	DNA (cytosine-5)-methyltransferase 1B-like	Y
V8896_e	2	20810259	LOC113731952	NADPH-dependent pterin aldehyde reductase	Y
V8905_e	2	21060957	LOC113731976	GTPase Der	NRN
V8905_c	2	21060957	LOC113726920	nucleolar GTP-binding protein 1-like	YRN
V9011_c	2	21195891	LOC113726937	NADH dehydrogenase [ubiquinone] iron-sulfur protein 8, mitochondrial	Y
V354_c	2	27616187	LOC113723028	subtilisin inhibitor CLSI-I-like	Y
V9189_c	2	31083032	LOC113727251	uncharacterized	Y
V9295_e	2	33094940	LOC113729006	cytochrome P450 87A3-like	Y
V147_e	2	53474348	LOC113729204	probable pectinesterase/pectinesterase inhibitor 25	Y

V9634_c	2	57478470	LOC113727543	uncharacterized	Y
V9697_c	2	58842880	LOC113727596	cation/calcium exchanger 5-like	Y
V9684_c	2	59202346	LOC113727613	cyclin-P3-1-like	Y
V9666_c	2	59551188	LOC113727629	DNA polymerase zeta catalytic subunit-like	Y
V9652_c	2	59888512	LOC113727657	heavy metal-associated isoprenylated plant protein 35-like	Y
V9802_c	2	60101536	LOC113727668	dormancy-associated protein homolog 3-like	Y
V9964_e	2	61762987	LOC113729328	protein FAR1-RELATED SEQUENCE 5-like	Y
V9932_c	2	62076582	LOC113727828	psbP domain-containing protein 1, chloroplastic-like	Y
V10155_c	2	64080402	LOC113728002	F-box protein PP2-A13-like	Y
V10177_e	2	64504213	LOC113732790	aspartic proteinase-like protein 2	Y
V10178_e	2	64504222	LOC113732790	aspartic proteinase-like protein 2	Y
V9708_e	2	64561829	LOC113732794	probable serine/threonine-protein kinase PBL7	Y
V9708_c	2	64561829	LOC113724697	protein O-glucosyltransferase 1-like	Y
V1360_c	3	3754775	LOC113734210	putative ion channel POLLUX-like 2	Y
V1360_e	3	3754775	LOC113736199	uncharacterized	Y
V491_e	3	11695179	LOC113736538	cellulose synthase-like protein G3	Y
V2036_e	3	15592880	LOC113737448	eIF-2-alpha kinase GCN2-like	Y
V2037_e	3	15592885	LOC113737448	eIF-2-alpha kinase GCN2-like	Y
V962_e	3	25515454	LOC113736734	calreticulin-3-like	Y
V1527_c	4	12805764	LOC113740321	COBRA-like protein 2	Y
V927_e	4	36103351	LOC113741010	DNA annealing helicase and endonuclease ZRANB3-like	Y
V213_e	5	18861837	LOC113743955	BTB/POZ domain-containing protein At5g48130-like	Y
V1314_e	5	25578972	LOC113687324	cysteine-rich receptor-like protein kinase 25	Y
V1315_e	5	25578998	LOC113687324	cysteine-rich receptor-like protein kinase 25	Y
V2357_c	5	36230080	LOC113690110	putative late blight resistance protein homolog R1C-3	Y
V1405_c	5	37804264	LOC113688755	uncharacterized	Y
V1981_e	6	18199439	LOC113697570	uncharacterized	Y

V1983_e	6	18199494	LOC113697570	uncharacterized	Y
V6318_c		26364086		pentatricopeptide repeat-containing protein	
	6		LOC113693594	At3g18110, chloroplastic-like	Y
V6319_c		26364122		pentatricopeptide repeat-containing protein	
	6		LOC113693594	At3g18110, chloroplastic-like	Y
V10038		54076570			
_c	6		LOC113692051	uncharacterized	Y
V856_c		2564680		probable UDP-arabinopyranose mutase 1	
	7		LOC113697895		Y
V856_e	7	2564680	LOC113701139	protein SHORT-ROOT-like	Y
V941_c		6028346		LOB domain-containing protein 41-like	
	7		LOC113698207		Y
V941_e		6028346		probable metal-nicotianamine transporter YSL7	
	7		LOC113701457		Y
V890_c	7	16431338	LOC113697853	uncharacterized	Y
V890_c	7	16431338	LOC113697851	protein LAZY 1-like	Y
V75_c		33706981		CCR4-NOT transcription complex subunit 11-like	
	7		LOC113699143		Y
V75_e	7	33706981	LOC113701244	F-box protein CPR1-like	Y
V1583_c		5251763		UDP-N-acetylglucosamine diphosphorylase 1-like	
	9		LOC113708815		Y
V1411_e		26900826		phospho-N-acetylmuramoyl-pentapeptide-transferase homolog	
	9		LOC113710463		Y
V1470_c	9	33200105	LOC113707872	uncharacterized	Y
V4918_e		2553024		probable galacturonosyltransferase-like 1	
	10		LOC113711862		Y
V4919_e		2553056		probable galacturonosyltransferase-like 1	
	10		LOC113711862		Y
V4920_e		2553077		probable galacturonosyltransferase-like 1	
	10		LOC113711862		Y
V4921_e		2594407		BEL1-like homeodomain protein 4	
	10		LOC113712363		Y
V4922_e		2594411		BEL1-like homeodomain protein 4	
	10		LOC113712363		Y
V5156_e	10	6590142	LOC113711742	uncharacterized	Y
V5184_e		7442209		XIAP-associated factor 1-like	
	10		LOC113710842		Y
V5963_c		7460043		putative 12-oxophytodienoate reductase 11	
	10		LOC113714564		Y

V5963_e	10	7460043	LOC113710944	probable rhamnogalacturonate lyase B	Y
V214_e	10	24056448	LOC113712412	beta-1,3-galactosyltransferase 7-like	Y
V215_e	10	24056497	LOC113712412	beta-1,3-galactosyltransferase 7-like	Y
V216_e	10	24056507	LOC113712412	beta-1,3-galactosyltransferase 7-like	Y
V217_e	10	24056512	LOC113712412	beta-1,3-galactosyltransferase 7-like	Y
V709_c	10	27791481	LOC113713621	nitrate reductase [NADH]-like	Y
V712_c	10	27791560	LOC113713621	nitrate reductase [NADH]-like	Y
V5514_e	10	34808941	LOC113712414	signal recognition particle 43 kDa protein, chloroplastic-like	Y
V5514_e	10	34808941	LOC113712415	replication factor C subunit 3	Y
V5361_e	10	36968487	LOC113712139	probable serine/threonine-protein kinase PBL12	Y
V220_e	10	38001494	LOC113712174	dehydration-responsive element-binding protein 1J-like	Y
V5699_c	10	40827076	LOC113713831	uncharacterized	Y
V5700_c	10	40827089	LOC113713831	uncharacterized	Y
V1286_c	10	42738120	LOC113714644	zinc finger CCCH domain-containing protein 55-like	Y
V5809_c	10	44037441	LOC113714295	abscisic stress-ripening protein 5-like	Y
V5810_c	10	44037450	LOC113714295	abscisic stress-ripening protein 5-like	Y
V5811_c	10	44037473	LOC113714295	abscisic stress-ripening protein 5-like	Y
V5838_c	10	44457920	LOC113713825	tryptophan aminotransferase-related protein 4-like	Y
V6038_c	11	4709133	LOC113717310	uncharacterized	Y
V6304_e	11	7623511	LOC113717847	uncharacterized	Y
V6305_e	11	7623569	LOC113717847	uncharacterized	Y
V6264_c	11	12674994	LOC113716958	serine/threonine-protein phosphatase 2A activator-like	Y
V6790_c	11	26648349	LOC113716118	probable S-adenosylmethionine-dependent methyltransferase At5g38100	Y
V6791_c	11	26648392	LOC113716118	probable S-adenosylmethionine-dependent methyltransferase At5g38100	Y

V6790_e	11	26648349	LOC113719108	probable disease resistance protein At1g58602	Y
V6791_e	11	26648392	LOC113719108	probable disease resistance protein At1g58602	Y
V6820_c	11	27900161	LOC113716177	probable inactive receptor kinase At5g16590	Y
V6880_c	11	28534901	LOC113717158	uncharacterized	Y
V6969_c	11	29381221	LOC113717020	uncharacterized	Y
V7006_c	11	29894818	LOC113717235	tryptophan synthase alpha chain-like	Y
V7053_c	11	30452457	LOC113716220	putative late blight resistance protein homolog R1A-10	Y
V7109_c	11	31449939	LOC113716538	serotonin N-acetyltransferase 2, chloroplastic	Y
V6594_c	11	32213390	LOC113716326	alpha-1,4 glucan phosphorylase L-2 isozyme, chloroplastic/amyloplastic-like	Y
V7191_c	11	32300796	LOC113715681	pentatricopeptide repeat-containing protein At2g03380, mitochondrial-like	Y
V6642_c	11	32681014	LOC113716244	myb-like protein AA	Y
V6690_c	11	33844511	LOC113717037	uncharacterized	Y
V6886_e	11	36615617	LOC113717766	probable sulfate transporter 3.3	Y
V6927_e	11	36901542	LOC113717916	putative late blight resistance protein homolog R1B-16	Y
V7409_e	11	40608337	LOC113718957	protein DETOXIFICATION 29-like	Y
V7438_e	11	41084720	LOC113717845	E3 ubiquitin-protein ligase AIRP2-like	Y
V7443_e	11	41135712	LOC113717564	DELLA protein GAI-like	Y
V7444_e	11	41135773	LOC113717564	DELLA protein GAI-like	Y
V7360_e	11	42289762	LOC113719459	transaldolase	Y

4 DISCUSSION

It is known that yield is a polygenic characteristic, thus, the study of dependencies between the characteristics that influence production at the level of molecular markers could be carried out, from the construction of a Bayesian Network, using the HC algorithm, which incorporates the four traits and finally being incorporated into a GWAS model based on SEM to decompose the SNP effects into direct and indirect on the trait.

Studies considering yield of *C. arabica* using GBLUP and ANOVA found low to medium heritability results, as seen in Sousa et al. (2019) and Weldemichael et al. (2017), who found a value of 0.26 and 0.28, respectively. Carvalho et al. (2019) considering the same method in *Coffea canephora* found 0.15, while in this study was found 0.14. For vegetative vigor, Sousa et al. (2019) found heritability of 0.34 and Carvalho et al. (2019) using GBLUP for canephora coffee found heritability of 0.23, while in this study, the heritability value was 0.39. For fruit size, in this study was found genomic heritability result of 0.61, while Sousa et al. (2019) obtained 0.36. For number of reproductive nodes in this study was found heritability result of 0.13, while Sousa et al. (2019) found 0.23. Thus, we can observe that the heritability values found are similar to those in the literature.

With the use of the Bayesian network together with SEM, it was possible to obtain coefficients of paths that interconnect important characteristics in Arabica coffee. It can be observed that there was a positive connection both directly (VV → Y) and indirectly (VV → NNR → Y) positive between VV and Y, thus indicating that the better the vegetative vigor status of the plant, the greater will be its production. Rodrigues et al. (2012), studying the influence of vegetative vigor on production, observed that it limits production. The opposite was observed when we analyzed the influence of NRN and FS on Y. Jaeggi et al. (2019) using path analysis, also has a negative relationship between NRN and Y. This relationship can be explained by the high drain for the development of many nodes, which leads to a reduction in the availability of nutrients for fruit formation.

Gene identification analysis based on information from the NCBI (2021) allowed detecting 189 SNP associated with the Yield, inserted in genes (Table 4). There were occurrences of markers allocated to the same gene on several chromosomes, as seen in the table 4. Their functional annotation and gene can be seen in same table. SNP that were not located within genes are also relevant for use as genetic markers in breeding. Molecular marker does not necessarily need to be inserted in a gene to detect genetic differences between individuals, it can be associated with the gene and be efficient (Andersen & Lubberstedt, 2013).

Furthermore, the SNP may be in promoter or regulatory regions, and therefore involved in gene expression.

According to the functional annotation of the genes in which the SNP are inserted, no mechanism was identified that has a direct influence on NRN control, however, for Y it was possible to identify some genes that have a direct influence on its control, such as the genes: i) LOC113731461_e - Stress enhanced protein 1, chloroplastic; ii) LOC113714295_c - Abscisic stress-ripening protein 5; iii) LOC113726102_c - Negative regulator of systemic acquired resistance SNI1. De acordo com Heddad & Adamska (1999), estudando *Arabdopsis thaliana*, the stress enhanced protein 1, chloroplastic pode desempenhar um papel fotoprotetor na membrana tilacóide em resposta ao estresse luminoso. Li et al, (2017), in rice studies, identified the involvement of Abscisic stress-ripening protein 5 in drought tolerance, 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), identified a negative reduction in gene expression and DNA recombination during a susceptible pathogen infection, therefore, involved in a short-term defense response and a long-term supply strategy.

5 CONCLUSION

With this study, it was possible to extend the study of the genome association study using several characters, adding a Bayesian network structure, and thus quantifying the genetic interrelationships between important characteristics of arabica coffee, so that it was possible to estimate the genetics direct and indirect effects and then understand the genetic architecture formed. Thus, we identified a positive interrelationship between vegetative vigor in production and vegetative vigor for the number of reproductive nodes and negative for the number of reproductive nodes and size of the fruit for production. It was also possible to detect significant genomic regions, and thus identify three genes that act directly on production.

6 REFERENCES

ABIC, Tendências do mercado de cafés em 2017. https://www.abic.com.br/wp-content/uploads/2020/01/2017.pdf

Brenner, E. A., Beavis, W. D., Andersen, J. R., Lübberstedt, T. 2013. Prospects and limitations for development and application of functional markers in plants. Diagnostics in Plant Breeding, p. 329-346.

Carvalho, H.F., Silva, F.L.D., Resende, M.D.V.D., Bhering, L.L. 2019. Selection and genetic parameters for interpopulation hybrids between kouilou and robusta coffee. Bragantia, 78, p. 52-59.

Cilas, C., Bar-Hen, A., Montagnon, C., Godin, C. 2006. Definition of architectural ideotypes for good yield capacity in Coffea canephora. Annals of Botany, 97(3), pp.405-411.

CONAB - ACOMPANHAMENTO DA SAFRA BRASILEIRA DE CAFÉ. Primeiro levantamento, janeiro de 2020. v. 6 - Safra 2020, n. 1.

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/

Durrant W.E., Wang S., Dong X. 2007. Arabidopsis SNI1 and RAD51D regulate both gene transcription and DNA recombination during the defense response. Proc Natl Acad Sci U S A. 2007 Mar 6;104(10):4223-7. doi: 10.1073/pnas.0609357104. Epub 2007 Feb 21. Erratum in: Proc Natl Acad Sci U S A. 104(17), p. 7307.

Ferrão, R. G.; Fonseca, A. F. A. Da.; Ferrão, M. A. G.; Verdin Filho, A. C.; Volpi, P. S.; De Muner. L. H.; Lani, J. A.; Prezotti, L. C.; Ventura, J. A.; Martins, D. Dos S.; Mauri, A. L.; Marques, E. M. G.; Zucateli, F. Café Conilon: Técnicas de Producion com variedades mejoradas. 2012. 4 ed. Revisada y actualizada, Vitória, ES: Incaper, 74p. (Circular Técnica, 03-I).

Ferrão, R.G., de Muner, L.H., da Fonseca, A.F.A., Ferrão, M.A.G. 2016. Café Conilon. Vitória, ES: Incaper, 2017.

Heddad, M., Adamska, I. 2000. Light stress-regulated two-helix proteins in Arabidopsis thaliana related to the chlorophyll a/b-binding gene family. Proceedings of the National Academy of Sciences, 97(7), p. 3741-3746.

Ikegawa, S. 2012. A short history of the genome-wide association study: where we were and where we are going. Genomics & informatics, 10(4), p. 220.

Jaeggi, M.E.P.C., Coelho, F.C., Pereira, I.M., Zacarias, A.J., de Amaral Gravina, G., de Lima, W.L., Pereira, L.L., Moreira, T.R., da Silva, S.F., do Carmo Parajara, M. 2019. Path analysis of vegetative characteristics in conilon coffee production consortiated with green fertilizers in tropical climate. Journal of Experimental Agriculture International, p. 1-11.

Jiang, G., Chakraborty, A., Wang, Z., Boustani, M., Liu, Y., Skaar, T., Li, L. 2013. New aQTL SNPs for the CYP2D6 identified by a novel mediation analysis of genome-wide SNP arrays, gene expression arrays, and CYP2D6 activity. BioMed research international, 2013.

Korb, K.B., Nicholson, A.E. 2010. Bayesian artificial intelligence. CRC press.

Korte, A., Vilhjálmsson, B.J., Segura, V., Platt, A., Long, Q., Nordborg, M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nature genetics, 44(9), p. 1066-1071.

Lashermes, P., Combes, M. C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., Charrier, A. 1999. Molecular characterisation and origin of the Coffea arabica L. genome. Molecular and General Genetics MGG, 261(2), p. 259-266.

Li J., Li Y., Yin Z., Jiang J., Zhang M., Guo X., Ye Z., Zhao Y., Xiong H., Zhang Z., Shao Y., Jiang C., Zhang H., An G., Paek N.C., Ali J., Li Z. 20167 OsASR5 enhances drought tolerance through a stomatal closure pathway associated with ABA and H2 O2 signalling in rice. Plant Biotechnol J. 15(2), p. 183-196.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., Borevitz, J.O. 2010. Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proceedings of the National Academy of Sciences, 107(49), p. 21199-21204.

Meuwissen, T.H., Hayes, B.J., Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), p. 1819-1829.

Meyer, K. 2007. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University Science B, 8(11), p. 815-821.

Meyer, K., Tier, B. 2012. “SNP Snappy”: a strategy for fast genome-wide association studies fitting a full mixed model. Genetics, 190(1), p. 275-277.

Mi, X., Eskridge, K., Wang, D., Baenziger, P.S., Campbell, B.T., Gill, K.S., Dweikat, I. 2010. Bayesian mixture structural equation modelling in multiple-trait QTL mapping. Genetics research, 92(3), p. 239-250.

Momen, M., Ayatollahi Mehrgardi, A., Amiri Roudbar, M., Kranis, A., Mercuri Pinto, R., Valente, B.D., Morota, G., Rosa, G.J., Gianola, D. 2018. Including phenotypic causal networks in genome-wide association studies using mixed effects structural equation models. Frontiers in genetics, 9, p. 455.

Momen, M., Campbell, M.T., Walia, H., Morota, G. 2019. Utilizing trait networks and structural equation models as tools to interpret multi-trait genome-wide association studies. Plant methods, 15(1), p. 1-14.

Momen, M., Campbell, M.T., Walia, H., Morota, G. 2019. Harnessing phenotypic networks and structural equation models to improve genome-wide association analysis. bioRxiv, p.553008.

O’Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R. and Coin, L.J., 2012. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS one, 7(5), p. e34861.

Rodrigues, M.N., Ferrao, R.G., Fonseca, A.F.A., Mendonca, R.L., Martins, L.D. and Tomaz, M.A., 2012. Crop yield of conilon coffee plants of different levels of vegetative vigor and rust severity. Nucleus, 9(2), p. 1-6.

Sant'Anna, G.C., Pereira, L.F., Pot, D., Ivamoto, S.T., Domingues, D.S., Ferreira, R.V., Pagiatto, N.F., da Silva, B.S., Nogueira, L.M., Kitzberger, C.S., Scholz, M.B., 2018. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. Scientific reports, 8(1), p. 1-12.

Scutari, M. 2009. Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817.

Sousa, T.V., Caixeta, E.T., Alkimim, E.R., Oliveira, A.C.B., Pereira, A.A., Sakiyama, N.S., Zambolim, L., Resende, M.D.V. 2019. Early selection enabled by the implementation of genomic selection in Coffea arabica breeding. Frontiers in plant science, 9, p. 1934.

Tran, H.T., Furtado, A., Vargas, C.A.C., Smyth, H., Lee, L.S., Henry, R. 2018. SNP in the Coffea arabica genome associated with coffee quality. Tree Genetics & Genomes, 14(5), p. 1-15.

USDA. 2021. Coffee: World Markets and Trade. https://apps.fas.usda.gov/psdonline/circulars/coffee.pdf

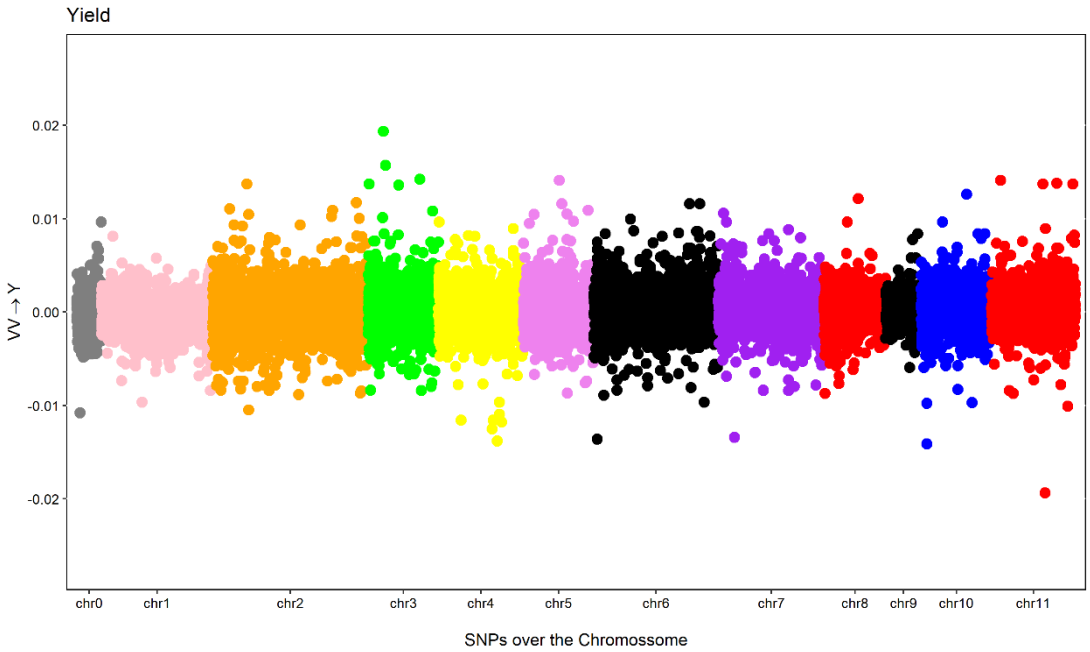
Valente B.D., Rosa G.J., Gianola D., Wu X-L, Weigel K.A. 2013. Is structural equation modeling advantageous for the genetic improvement of multiple traits?. Genetics, 194(3), p. 561–572.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), pp.4414-4423.

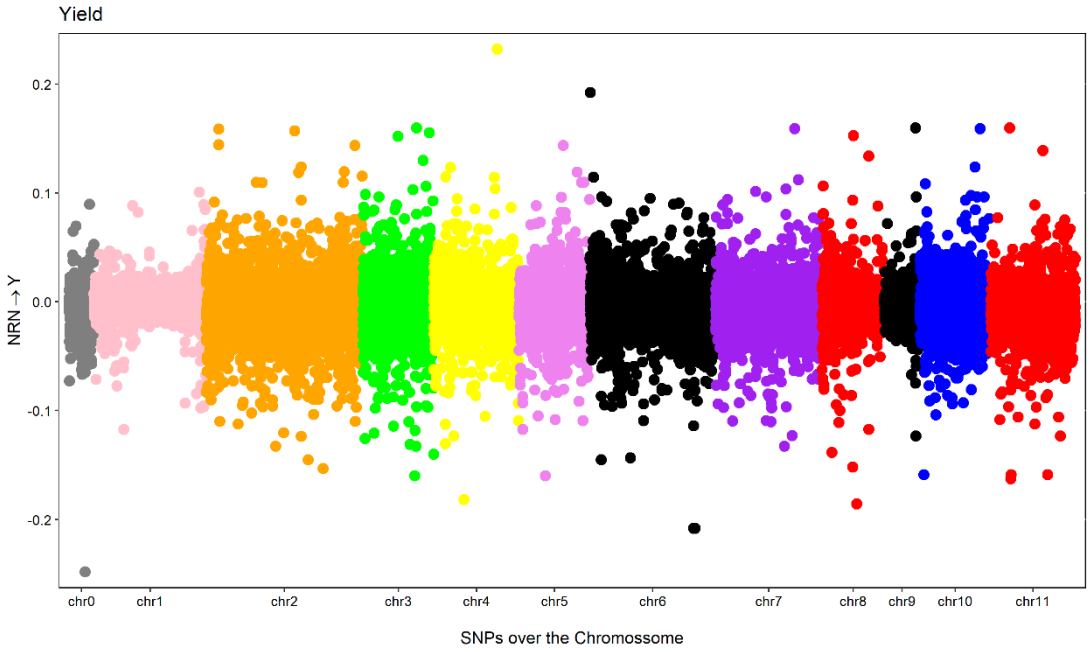
Weldemichael, G., Alamerew, S., Kufa, T. 2017. Genetic variability, heritability and genetic advance for quantitative traits in coffee (Coffea arabica L.) accessions in Ethiopia. African Journal of Agricultural Research, 12(21), pp.1824-1831.

Zhou, X., Stephens, M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nature genetics, 44(7), p.821.

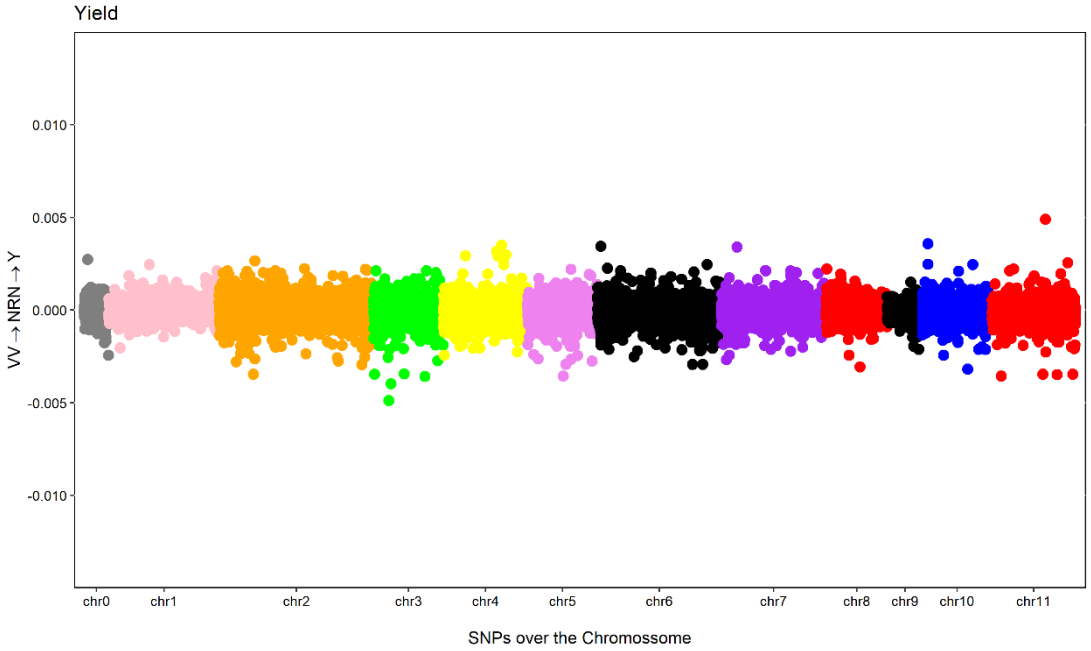
7 ATTACHMENTS



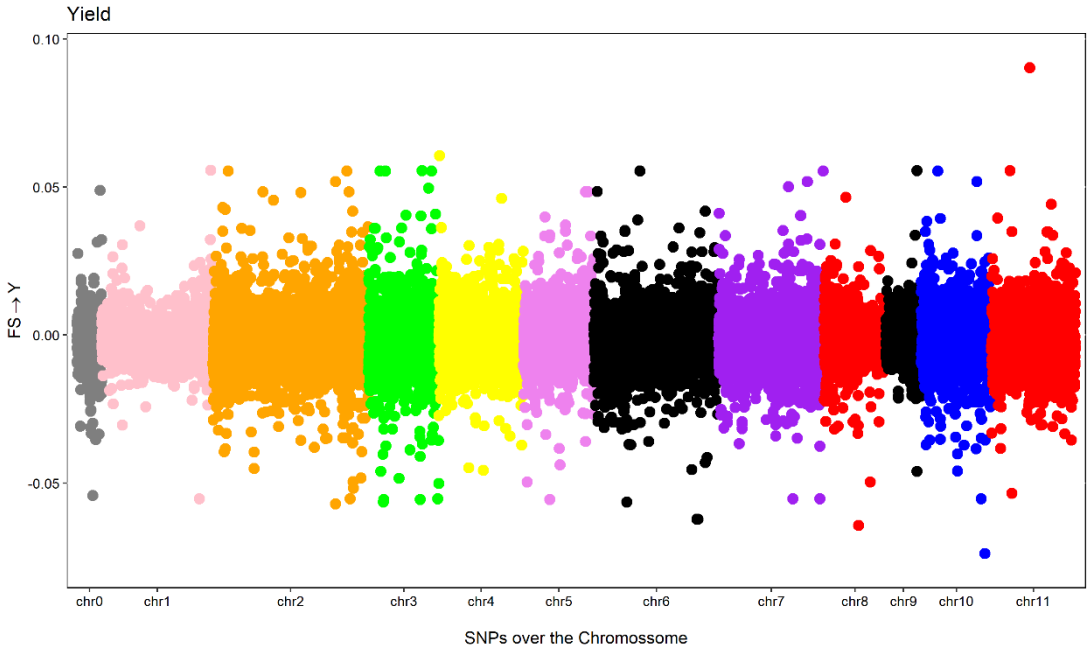
SF1: Manhattan plots with enlargement for SNP effects on Yield obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. VV: vegetative vigor; Y: Yield.



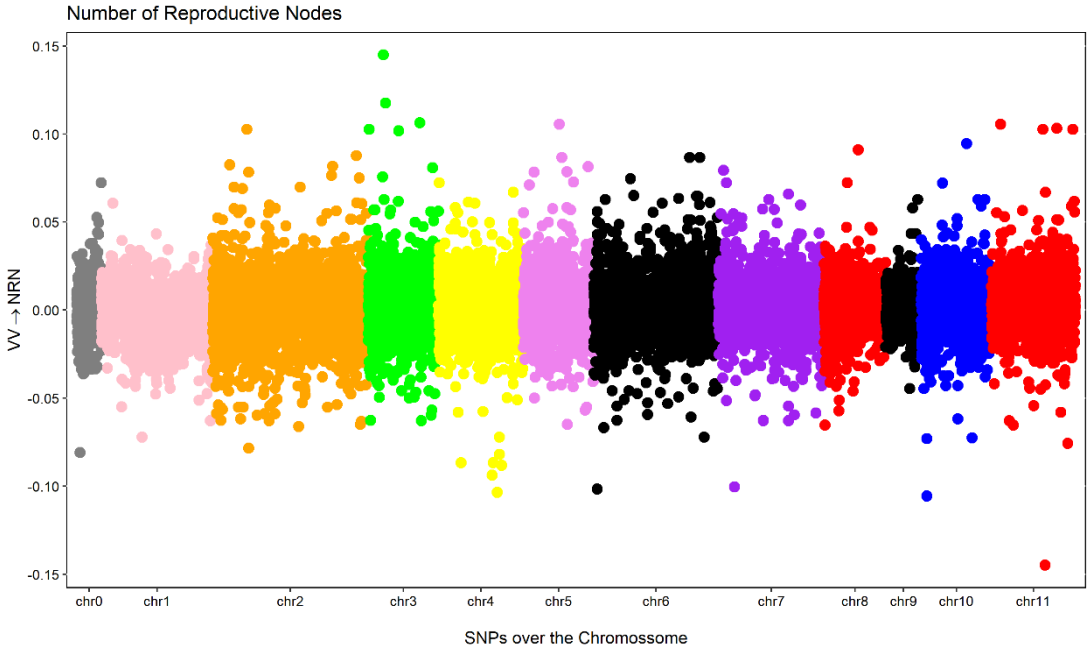
SF2: Manhattan plots with enlargement for SNP effects on Yield obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. NRN: number of reproductive nodes; Y: Yield.



SF3: Manhattan plots with enlargement for SNP effects on Yield obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. VV: vegetative vigor; NRN: number of reproductive nodes; Y: Yield.



SF4: Manhattan plots with enlargement for SNP effects on Yield obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. FS: fruit size; Y: Yield.



SF5: Manhattan plots with enlargement for SNP effects on number of reproductive nodes obtained using SEM-GWAS based on the network structure learned by Hill Climbing algorithm. VV: vegetative vigor; NRN: number of reproductive nodes.

Convergence analysis

#niter=1.2M; burnin=50k; thin=50
R

LAGS AND AUTOCORRELATIONS:
==========

Chain: R

	Lag 1	Lag 5	Lag 10	Lag 50
V1	0.04149533	0.0235427939	-0.006817042	0.0033779687
V2	0.02961728	0.0063482279	-0.007710827	0.0117975961
V3	0.03231226	0.0145107318	-0.004858921	-0.0016827318
V4	0.09297639	0.0273947132	-0.006736830	-0.0092522235
V5	0.01762976	0.0110801337	-0.005240808	0.0076986133
V6	0.03700064	0.0205398452	0.008732112	0.0116028048
V7	0.05790439	0.0006186543	-0.012360439	0.0044281055
V8	0.03237568	-0.0009093089	-0.005749035	-0.0006198499
V9	0.06379039	0.0023251174	-0.013951965	-0.0093901876
V10	0.13238355	0.0119532470	0.001860423	-0.0076808849

GEWEKE CONVERGENCE DIAGNOSTIC:
==========

Fraction in first window = 0.1
Fraction in last window = 0.5

Chain: R

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Z-Score	-1.1294541	0.5574693	0.8888514	-0.2998819	2.3666849	2.13349019	0.1313820	-0.3700722	-0.7448091	0.817195
p-value	0.2587063	0.5772068	0.3740830	0.7642673	0.0179482	0.03288454	0.8954731	0.7113287	0.4563871	0.413817

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:
==========

Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

Chain: R

	Thin	Burn-in	Total	Lower Bound	Dependence Factor
V1	1	2	3945	3746	1.053123
V2	1	2	3685	3746	0.983716
V3	1	2	3918	3746	1.045916
V4	1	3	4083	3746	1.089963
V5	1	2	3761	3746	1.004004
V6	1	1	3747	3746	1.000267
V7	1	2	3813	3746	1.017886
V8	1	2	3787	3746	1.010945
V9	1	2	3839	3746	1.024826
V10	1	3	4126	3746	1.101442

G

LAGS AND AUTOCORRELATIONS:
==========

Chain: G

	Lag 1	Lag 5	Lag 10	Lag 50
V1	0.50239132	0.077037131	0.0071986248	-0.0179243518
V2	0.28350716	0.002370659	-0.0137969840	-0.0026106310
V3	0.25079798	0.006981694	0.0006316473	0.0096776876
V4	0.49937999	0.056537403	0.0122162507	-0.0064765711
V5	0.08125254	-0.007490150	-0.0010329143	-0.0107218453

V6 0.10319785 0.009822392 -0.0043233122 -0.0018757567
V7 0.19415597 0.008209455 -0.0043513509 -0.0043207583
V8 0.08347748 0.009255031 -0.0051651627 0.0008708542
V9 0.18075899 -0.006594235 0.0005265323 0.0035707331
V10 0.30557140 0.006918642 0.0091229355 0.0015180381

GEWEKE CONVERGENCE DIAGNOSTIC:
==========

Fraction in first window = 0.1
Fraction in last window = 0.5

Chain: G

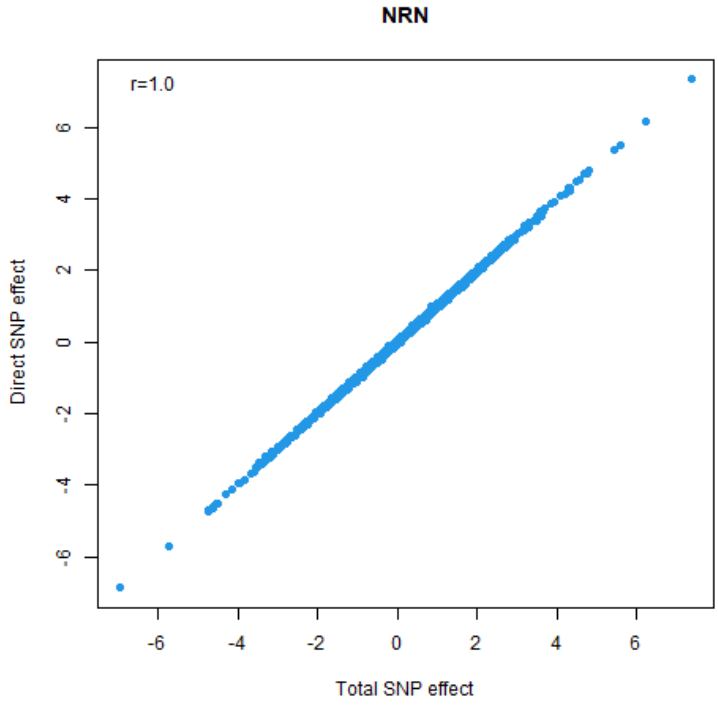
V1 V2 V3 V4 V5 V6 V7 V8 V9
Z-Score 0.2529986 0.2349770 0.02624938 0.2208149 -1.78450922 -1.1372147 1.73556554 -0.4125141 1.6217126
p-value 0.8002693 0.8142266 0.97905843 0.8252366 0.07434096 0.2554485 0.08264068 0.6799627 0.1048649
V10
Z-Score -0.2016144
p-value 0.8402182

RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:
==========

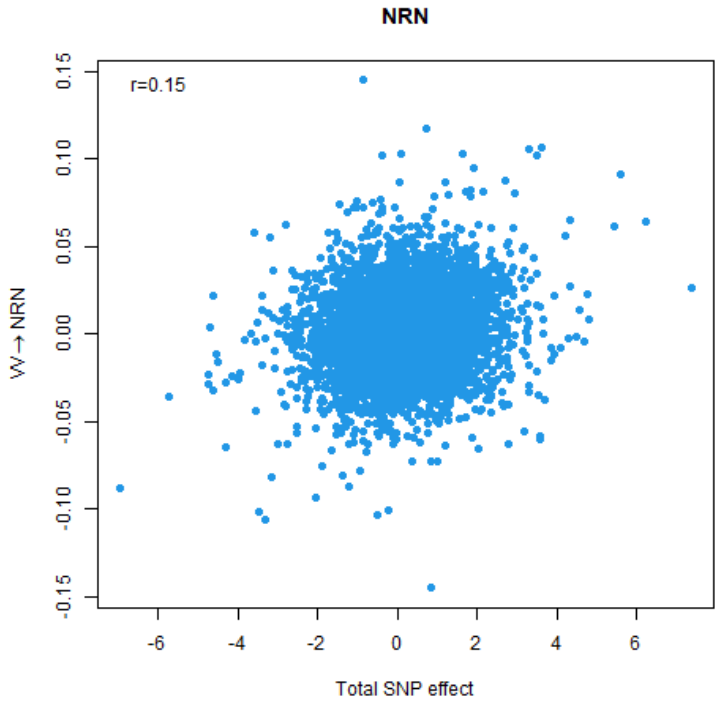
Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

Chain: G

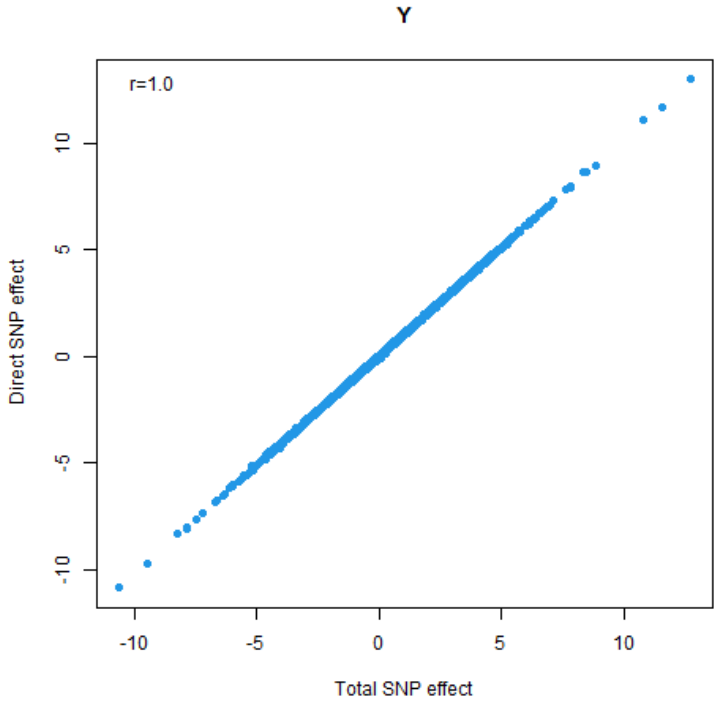
	Thin	Burn-in	Total	Lower Bound	Dependence Factor
V1	1	3	4434	3746	1.183663
V2	1	4	4638	3746	1.238121
V3	1	3	4306	3746	1.149493
V4	1	4	4719	3746	1.259744
V5	1	2	3905	3746	1.042445
V6	1	2	3774	3746	1.007475
V7	1	2	3972	3746	1.060331
V8	1	2	3865	3746	1.031767
V9	1	3	4028	3746	1.075280
V10	1	3	4403	3746	1.175387



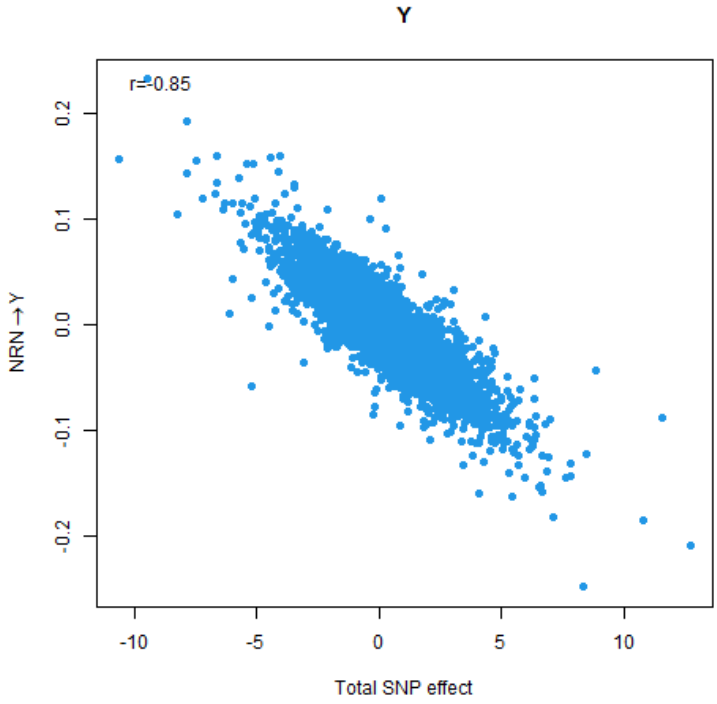
SF6: Correlation plots of decomposed SNP effects for NRN. Each point corresponds to the estimated effect of a SNP which direct affects NRN.



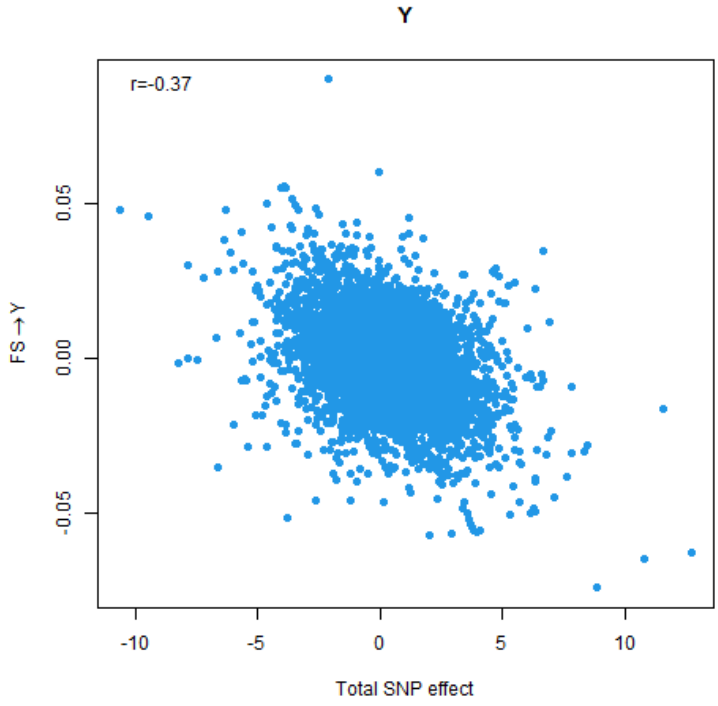
SF7: Correlation plots of decomposed SNP effects for NRN. Each point corresponds to the estimated effect of a SNP which indirect affects NRN.



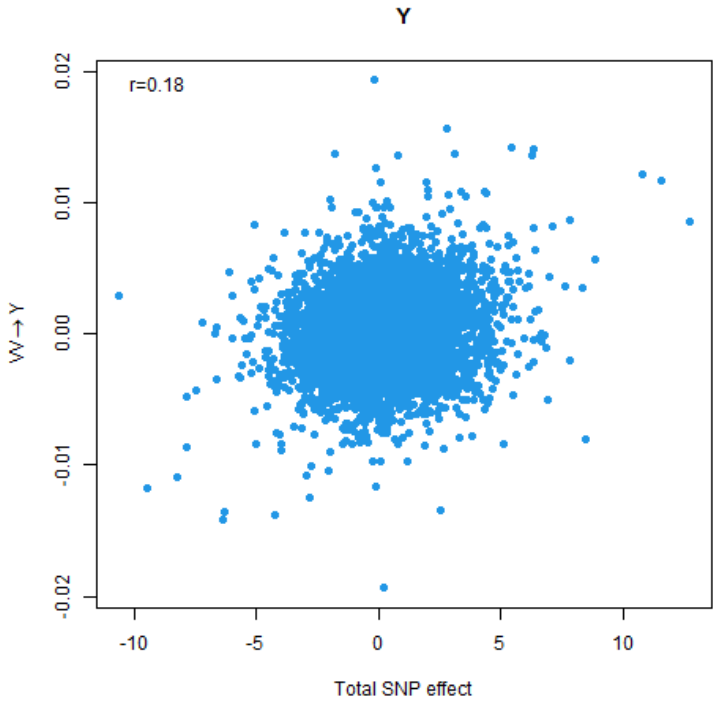
SF8: Correlation plots of decomposed SNP effects for Y. Each point corresponds to the estimated effect of a SNP which direct affects Y.



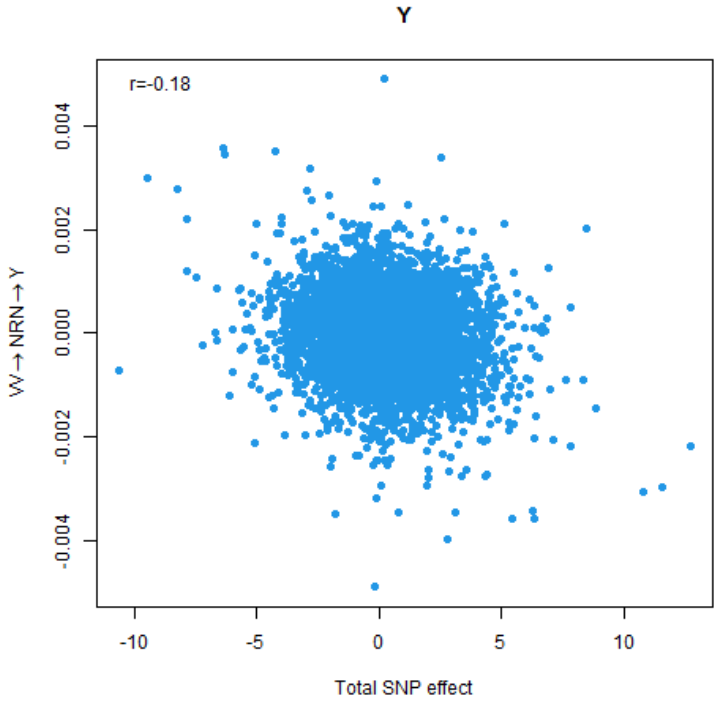
SF9: Correlation plots of decomposed SNP effects for Y. Each point corresponds to the estimated effect of a SNP which indirect affects Y.



SF10: Correlation plots of decomposed SNP effects for Y. Each point corresponds to the estimated effect of a SNP which indirect affects Y.



SF11: Correlation plots of decomposed SNP effects for Y. Each point corresponds to the estimated effect of a SNP which indirect affects Y.



SF12: Correlation plots of decomposed SNP effects for Y. Each point corresponds to the estimated effect of a SNP which indirect affects Y.