Euphytica (2021) 217:190
https://doi.org/10.1007/s10681-021-02922-9

Association mapping reveals genomic regions associated with bienniality and resistance to biotic stresses in arabica coffee

Juliana Vieira Almeida Nonato · Humberto Fanelli Carvalho · Karina Lima Reis Borges · Lilian Padilha · Mirian Perez Maluf · Roberto Fritsche-Neto · Oliveiro Guerreiro Filho

Received: 25 March 2021 / Accepted: 6 September 2021 / Published online: 23 September 2021

Abstract The bienniality of production and the incidence of pests and diseases, such as coffee leaf miner and coffee leaf rust, stands out among the factors that limit coffee crop yield. Obtaining cultivars with greater stability in production and resistance to these biotic agents are among the main objectives of coffee breeding programs. In this way, biotechnological tools such as Genomic Wide Association Studies (GWAS) can increase these programs' efficacy since they allow the identification of molecular markers significantly associated with phenotypes of interest. In this context, the aim here is to identify genomic regions associated with yield, bienniality, and resistance to coffee leaf miner and coffee leaf rust in arabica coffee progenies. Thus, a population (n = 597) was evaluated for resistance to biotic stresses and for the eight designed scenarios to study yield and bienniality. A matrix of 4,666 SNPs (Single Nucleotide Polymorphism) was built through Genotyping by Sequencing (GBS). After the genomic association analyses, we identified 12 potential SNPs markers associated with resistance to coffee leaf miner and coffee leaf rust, 32 associated with the eight designed scenarios to study yield and bienniality. Of the 44 SNPs significantly associated with this study's traits, 36 were noted in genomic regions responsible for biological processes related to plant response to biotic and abiotic stresses. In addition, four markers were coincident with yield and traits related to coffee leaf rust resistance. The genomic regions identified in this study can be incorporated into the coffee breeding program, through assisted selection, leading to more efficient breeding strategies in coffee.

Keywords GWAS · Coffee leaf rust · Coffee leaf miner · Coffee resistance · SNP markers

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10681-021-02922-9.

J. V. A. Nonato · H. F. Carvalho · O. Guerreiro Filho (✉)
Agronomic Institute of Campinas (IAC), Coffee Center, Campinas, São Paulo, Brazil
e-mail: oliveiro.guerreiro@sp.gov.br

K. L. R. Borges · R. Fritsche-Neto
Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, São Paulo, Brazil

L. Padilha · M. P. Maluf
Brazilian Agriculture Research Corporation (EMBRAPA), Coffee, Campinas, São Paulo, Brazil

Introduction

Coffee farming represents an important agricultural activity worldwide. In terms of economic relevance *Coffea arabica* L. species stands out due to the high

quality of its beverage, being responsible for about 60% of world coffee production (ICO 2020). Despite the great importance of the crop, the incidence of pests and diseases such as coffee leaf miner, *Leucoptera coffeella* (Dantas et al. 2020) and coffee leaf rust, *Hemileia vastatrix* (Pereira et al. 2021) affect crop productivity s in Brazil. Bienniality is other aspect affecting arabica coffee productive potential, which consists of alternating between low and high productivity in consecutive years, difficulting to estimate the average production of crops (Rena and Maestri 1987; DaMatta et al. 2007; Volsi et al. 2019; Carvalho et al. 2020).

The species *C. arabica* is allotetraploid (2n = 4x = 44) originated from the natural crossing between the diploid species *C. canephora* and *C. eugenioides*. However, it consists of two sub-genomes and behaves as a functional diploid (Lashermes et al. 2009, 2014). The genetic basis of *C. arabica* is relatively narrow due to the autogamous nature of the species, the small number of seeds or plants used in its commercial dispersion, and the frequent use of the genealogical method to select new cultivars (Setotaw et al. 2013). Besides the low variability among commercial cultivars of arabica coffee, other aspects such as the perennial cycle, the long juvenile period, and the phenotypic expression over several years make it difficult to apply classical improvement techniques. The use of molecular tools by breeding programs is an alternative to increase efficiency and reduce time and cost along selection cycles of superior genotypes (Ceccarelli 2015).

Next-generation sequencing (NGS) technologies have been used for whole genome sequencing of several crops, discovering large numbers of single nucleotide polymorphisms (SNPs) for exploration of within-species diversity, construction of haplotype maps and execution of Genome-Wide Association Studies (GWAS) (Nguyen et al. 2019). Recently, GWAS has been incorporated into breeding programs and this approach can identify the closest link between marker and phenotype, and it can be used to understand the genetic architecture of complex traits (Yang et al. 2013; Richards et al. 2017). The main advantage of the GWAS is that the analyzes could be performed with different types of populations, such as germplasm collections, cultivars or breeding lines. In these populations, more recombination events are observed, which allows for a higher resolution compared to linkage mapping (Tibbs Cortes et al. 2021). In addition, *C. arabica* low polymorphism makes the construction of genetic maps difficult (Pestana et al. 2015). The advance of sequencing technologies reduced the time and cost of DNA sequencing, allowing the production of a large number of genomic data. Genotyping by Sequencing (GBS) is a simple and robust methodology that uses NGS producing high-quality SNPs at a low cost per data point (Elshire et al. 2011).

Recent studies in *C. arabica* used SNPs in studies of linkage map (Moncada et al. 2016), genetic diversity (Sousa et al. 2017), and genomic prediction (Sousa et al. 2019; Carvalho et al. 2020). In addition, several genomic association studies conducted in *C. arabica* identified genomic regions associated with lipid and diterpene levels (Sant'Ana et al. 2018), 40 genes involved in the production, transport, and metabolism of caffeine, and 24 associated with trigonelline (Tran et al. 2018) and markers SNPs significantly associated with resistance to the coffee berry disease (CBD) (Gimase et al. 2020).

Association mapping is a promising genomic tool to detect associations between molecular markers and traits of interest (Zhang et al. 2018). The main advantage of applying GWAS in perennial crop breeding programs such as coffee is that it does not require the formation of a bi-parent population, reducing time and costs. So far, several studies of genomic association have been conducted on perennial species such as citrus (Minamikawa et al. 2017; Imai et al. 2018), cocoa (Romero Navarro et al. 2017), eucalyptus (Kainer et al. 2019), and coffee (Tran et al. 2018; Sant'Ana et al. 2018; Gimase et al. 2020). Thus, the present work intends to identify molecular markers (SNPs) associated with genomic regions responsible for the bienniality, resistance to the coffee leaf miner, and coffee leaf rust in *C. arabica* progenies.

Material and methods

Genetic material and phenotyping

The population used in this study consists of a progeny trial installed in 2003 at the Experimental Center of Agronomic Institute (IAC), SP State, Brazil (22°51'S, 47°04'W, altitude 640 m). The progeny trial was designed in a randomized complete block with ten

treatments, represented by nine segregating progenies (P_2 to P_{10}) and one experimental control (P_1—Catuaí Vermelho IAC 99 cultivar), nine blocks and a random number of plants by plot, a total of 599 plants of *C. arabica*. The nine progenies are from self- or open pollination of selected coffee trees (Supplementary Table S1). They segregate for resistance to the coffee leaf miner due to *C. racemosa* in the initial crosses with *C. arabica,* and for coffee leaf rust resistance due to introgression of genes of *C. canephora* and the germplasm Icatu (Supplementary Fig. S1).

The phenotyping of the population under study occurred along the years 2005 and 2012 and the following traits were evaluated by plant:

1. Yield (Y): evaluation carried out between 2005 and 2008 by weighing, in grams, fruits at different stages of maturity.
2. Resistance to the coffee leaf miner: evaluated under natural infection conditions for two consecutive years (2011 and 2012), in periods of high population intensity of the insect, which generally occur in April/May (first evaluation—LM1) and September/October (second evaluation—LM2). The periods in which the evaluations were carried out corresponded to different phenological stages of the plant. Therefore, they were considered as two distinct traits. The evaluations were carried out using a scale of notes ranging from 1 to 5 according to *L. coffeella*'s attack's intensity, being the plants classified with 1 as little infested (resistant plants) and those with 5 with severely infested (susceptible plants) (Guerreiro Filho et al. 1999).
3. Resistance to the coffee leaf rust: evaluated under natural infection conditions for two consecutive years (2011 and 2012), in periods of the high incidence of the disease, through the reaction type (LRRT—Leaf Rust according to the Reaction Type) and the lesions density (LRLD—Leaf Rust according to the Lesion Density—LRLD) in plants. Reaction type (LRRT) was evaluated using a scale of notes ranging from 1 to 5, where 1—plant immune without lesions and 5—susceptible plant presenting generalized pustules with many spores and severe defoliation. Lesion density was evaluated to observe the incidence of coffee leaf rust using the entire plant as a reading unit. LRLD were also evaluated according to a 1–10 scale of notes, where 1—absence of sporulating lesions on the plant, 2—occurrence of few lesions in branches of lower third of the plant, 3 to 9—gradual increase in the number of diseased branches on the plant, and 10—high incidence of disease, lesions in almost all branches, from the base to the apex of the plant (Eskes and Braghini 1981).

Phenotypic analysis

The phenotypic data analysis for the characters related to resistance to the coffee leaf miner and coffee leaf rust was performed through the MCMCglmm (Hadfield 2010) R package. The model considered was:

$$y = Xa + Wb + Tg + Zu + e \quad (1)$$

where y is the observation vector; a is the year effect vector, considered as fixed; b is the block effect vector, considered as fixed; g is the genotype effect vector, considered as random, where $g \sim N(0, I\sigma_g^2)$; u is the effect vector of year x genotype interaction, considered as random, where $u \sim N(0.I\sigma_u^2)$; *and* e is the random effect of residual, with $e \sim N(0, I\sigma^2)$. X, W, T, and Z are the incidence matrices of the respective effects. The genotypic values (BLUPs) obtained through this model were used to obtain Pearson's correlations between traits and estimate broad-sense heritability.

Additionally, the adjusted phenotypic mean, BLUEs (*Best Linear Unbiased Estimator*) for yield (Y) was obtained for each year individually through the equation of the model described below:

$$y_{ik} = \mu + bk + gi + eik \quad (2)$$

where *yik* is the phenotypic value for the genotype i and the block k; μ is the average of the population; *bk* is the fixed effect of the k-th block; *gi* is the fixed effect of the i-th genotype; and *eik* is the random effect of the residues, where $e \sim N(0, I\sigma_e^2)$.

We have also designed eight different scenarios for the Y variable in order to investigate the production and bienniality (BIEN) over the years evaluated, similar to Carvalho et al. (2020):

1. Scenario 1: analysis conducted for the four consecutive years (2005, 2006, 2007, and 2008). This scenario simulates the actual situation of the

coffee breeding program, in which the production of the plants is evaluated individually for at least four consecutive harvests;

2. Scenarios 2, 3, 4, and 5: analyses performed for each of the four years individually (2005, 2006, 2007, and 2008), respectively. These scenarios were designed in order to detect SNPs that could be useful to identify productive genotypes in several years;
3. Scenario 6: analyses performed with the estimates of MHPRVG (Harmonic Average of Relative Performance of Genetic Values) for the first biennium between the consecutive years 2005 and 2006. It simulates the effect of the biennial production;
4. Scenario 7: analyses performed with MHPRVG estimates for the second biennium between consecutive years 2007 and 2008. Same scenario 6;
5. Scenario 8: analyses performed with arithmetic mean between scenarios 6 and 7.
6. For scenarios 6, 7, and 8, described above, the MHPRVG, proposed by Resende (2002), was estimated through the biennial combination between consecutive years (2005/2006 and 2007/2008) and calculated by following equation:

$$MHPRGV = \frac{n}{\sum_{j=1}^{n}\left(\frac{GV_{ij}}{M_j}\right)^{-1}} \quad (3)$$

where n is the number of genotypes, $_{Mj}$ is the average of the year, and $_{GVij}$ is the genotypes' genetic values.

Genotyping of SNP markers

Total genomic DNA was extracted from leaves collected of each individual plant, using the CTAB protocol (Doyle and Doyle 1990). The libraries' construction was carried out according to the protocol of Elshire et al. (2011), using the restriction enzyme *Pst1*. The libraries were sequenced on Illumina-HighSeq 2500 platform with 48 samples per sequencing *lane*. The sequences resulting from the sequencing were aligned in the reference genome of *C. arabica* (https://www.ncbi.nlm.nih.gov/genome/?term=txid13443[orgn]]) through Bowtie 2 (Langmead and Salzberg 2012). The SNP calling and imputation of missing data were performed in FreeBayes (Garrison and Marth 2012) and BEAGLE 5.0 (Browning et al. 2018). The SNPs, in turn, were filtered through VCFtools (Danecek et al. 2011), in which only biallelic markers were maintained, and the SNPs with allele frequencies below 5% MAF (> 0.05) and call rate (> 0.80) were removed. Finally, the SNPs were submitted to the linkage disequilibrium (LD) filter, in which the combinations in pairs with values higher than 0.99 were removed.

Analysis of the population structure and the linkage disequilibrium

The population structure (Q) was evaluated through the principal component analysis (PCA), and the genomic kinship matrix (K) was obtained based on the frequencies of the alleles observed through the equation proposed by VanRaden (2008). The LD was estimated through the correlation coefficient between the alleles (r^2). The r^2 values were calculated for all pairs of SNP markers on each chromosome through the *synbreed* package (Wimmer et al. 2012) and the graphics of the LD-decay were generated in R package *ggplot2* (Wickham, 2016).

Genome-Wide association study

The GWAS was performed using the Mixed Linear Models method, through the FarmCPU package of the R software (Liu et al. 2016), and the following statistical model is described below:

$$y = Xb + Qp + Ku + e \quad (4)$$

where y is the vector of adjusted means of the genotypes, b is a vector with the fixed effects of the SNP markers; p is the vector of fixed effect of the population structure (PC); u is the random effect of relative *kinship* (*kinship* matrix), where $u \sim N(0, K\sigma_u^2)$; e is the random residual effect vector, where $e \sim N(0,\sigma_e^2)$. X, Q, and Z are the incidence matrices for these effects.

The dispersion of associations between SNP markers and the trait of interest was observed through Manhattan plots. The quality of the associated SNPs and the model's adjustment were verified in Quantile–Quantile plots (Q-Q plots), in which the associations found and expected are listed. The p-value of each SNP and its significance was defined assuming the

Bonferroni correction threshold, where $p < 0.05/n$, being n = total number of SNPs.

Identification of candidate genes

The reference genome of the species *C. arabica* (https://www.ncbi.nlm.nih.gov/genome/?term=txid13443[orgn]]) was used in the identification and analysis of candidate genes. The search for coding regions was carried out with the flanking sequences (5000 bp upstream and 5000 bp downstream) of the SNPs identified as significant for the traits analyzed in this study. This step was performed through annotation of genes based on sequences available at the National Center for Biotechnology Information (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using the nucleotide database (BLASTn). We selected only the regions with the highest similarity and lowest e-value.

Results

Phenotypic responses

Figure 1 summarizes phenotypic data distribution regarding yield, resistance to coffee leaf miner, and coffee leaf rust. The results related to yield indicate the occurrence of a biennial cycle in the population, in which the years 2005 and 2007 were of low yield and the years 2006 and 2008 of high production (Fig. 1a).

The evaluation of response to coffee leaf miner (Leaf Miner in April/May—LM1 and Leaf Miner in September/October—LM2) and to coffee leaf rust (Leaf Rust according to the Reaction Type—LRRT and Leaf Rust according to the Lesion Density—LRLD) shows that all progenies still segregate for the resistance trait to the two biotic agents. However, the frequency of resistant and susceptible plants varies from one progeny to another, although always with susceptible individuals predominance. For insect resistance, progenies 2, 5, 8, and 10 presented genotypes with a lower level of infestation (Fig. 1b and c). For coffee leaf rust resistance (LRRT and LRLD), all the progenies presented plants with a higher resistance level than that observed in susceptible control (progeny 1), as shown on Fig. 1d and e. The results presented in Fig. 1 points to a biannual behavior of production. A differential reaction refers to the average level of resistance to the coffee leaf miner and coffee leaf rust among the progenies, which shows the variability present in the population under study.

In addition, we estimated the heritability and phenotypic correlation between the traits. We observed a low and non-significant correlation between yield (Y) and the resistance to coffee leaf miner (LM1 and LM2) and coffee leaf rust (LRRT and LRLD) traits. However, considering the correlation only between the variables related to resistance (LM1, LM2, LRRT and LRLD), the values were positive and significant (Table 1). The estimated heritability value in the broad sense for the production trait was low (0.23), as expected since this trait is considered quantitative (Table 1). However, the observed values were higher for traits LM1, LRRT, and LRLD (> 0.90), except for LM2, which, among the resistance traits, had the lowest value (0.49).

Genomic data and population structure

A total of 597 plants were sequenced by GBS and the reads aligned with the reference genome of *C. arabica* resulted in an initial matrix of 120,617 SNP markers, distributed over the 22 chromosomes of *C. arabica*. Aiming to obtain a greater data reliability for the GWAS analysis, these markers were submitted to quality control (MAF > 0.05; call rate > 0.80) and, after keeping only the biallelic markers, we identified 8108 SNPs. Further, these SNPs were subjected to LD pruning, in which paired combinations with values greater than 0.99 were removed, resulting in a matrix of 4666 informative SNPs. In the end, this matrix was used to calculate the LD decay and was also used in genomic association analyses.

The LD decay graph was plotted with the squared correlation coefficient (r^2) and the physical distance using a non-linear regression. We observed a relatively slow LD decay in a window of 1000 kb (Supplementary S2). This regression curve pattern showed that the LD decayed at a distance of 48 kb ($r^2 = 0.11$). Extent of LD can be influenced by several factors, such as a small number of individuals in the population, low recombination rate, and the type of species reproduction (Gupta et al. 2005; Vos et al. 2017). Thus, LD decay behavior was expected in *C. arabica* due to the autogamous nature of this species. Similar results have been reported in studies of others

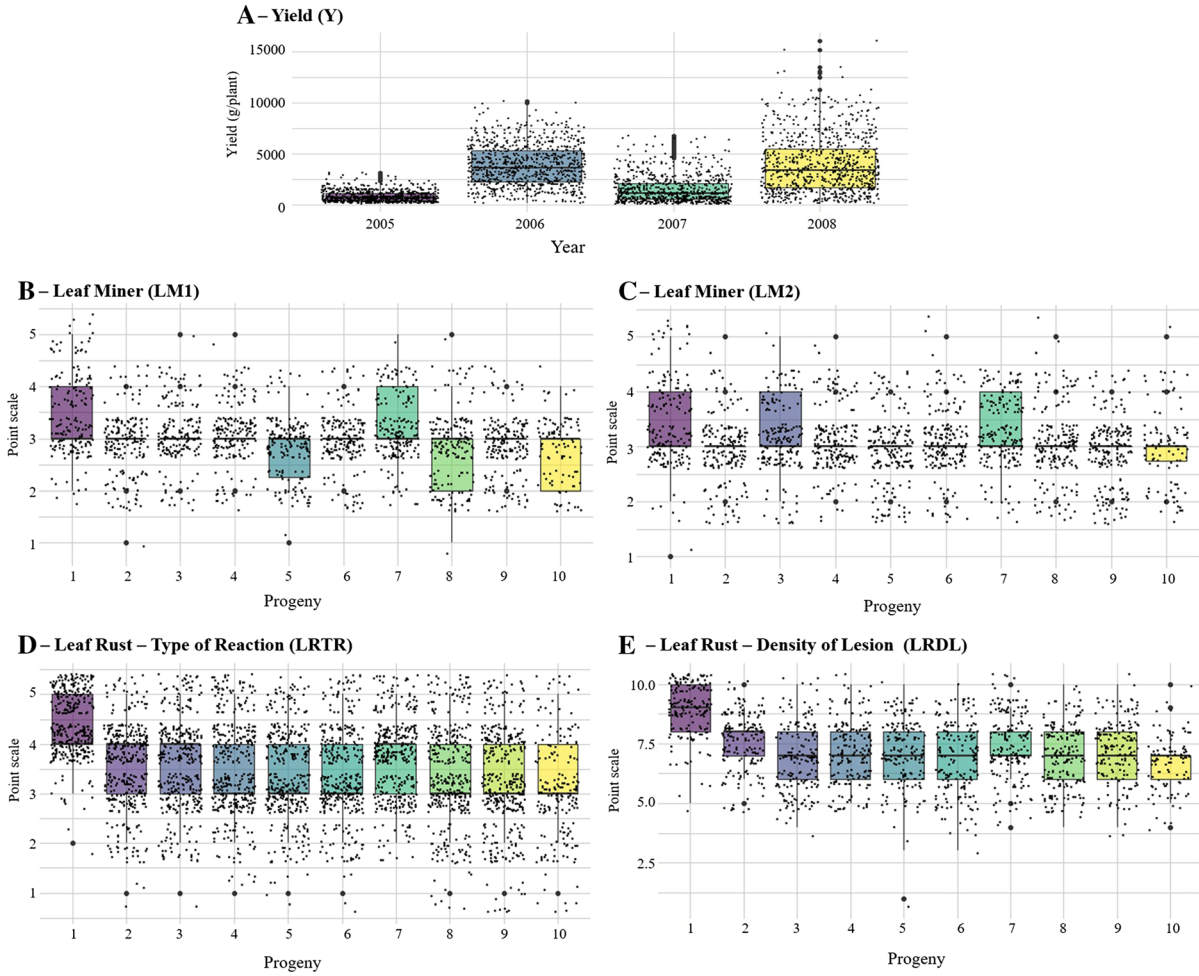


Fig. 1 Distribution of raw phenotypic data, in boxplot format, of the five traits evaluated in Arabica coffee progenies. **a** corresponds to yield, **b** first evaluation of resistance to the coffee leaf miner (LM1), **c** second evaluation of resistance to the coffee leaf miner (LM2), **d** resistance to coffee leaf rust according to the reaction type—LRRT, and **e** resistance to coffee leaf rust according to the lesion density (LRLD)

autogamous species such as sunflower and linseed (Kolkman et al. 2007; Singh et al. 2016).

We explored graphically the population structure of SNP markers using two methods (Fig. 2). The genetic structure of the genotypes was verified from the principal component analysis (PCA), whose principal components 1 (PC1) and 2 (PC2) explained, respectively, 54% and 17% of the variance and showed, in plan 1/2 (Fig. 2a), the dispersion of the population in six subgroups. Likewise the PCA, the genomic kinship matrix (K), visualized through *Heatmap*, indicated a genomic relationship between the different genotypes evaluated (Fig. 2b). These two parameters (matrix Q, obtained through the PCA and matrix K) were incorporated in the GWAS analyses in order to correct possible spurious associations that may occur due to the kinship relationship.

GWAS analyses and candidate genes associated to yield, bienniality and resistance of *C. arabica* to biotic stresses

The mixed linear model was used for the GWAS analyses in which the number of principal components that compose the Q matrix varied according to the model setting (Q-Q plots) of each analyzed trait. A

Table 1 Heritability in the broad sense and Pearson's correlation between yield (Y), resistance to the coffee leaf miner (LM1 and LM2), and coffee leaf rust evaluated by the reaction type (LRRT), and lesion density (LRLD) in the progenies of Arabica coffee

Trait	h^2	Correlation				
		Y	LM1	LM2	LRRT	LRLD
Y	0.23	–				
LM1	0.98	− 0.02 ns	–			
LM2	0.49	− 0.03 ns	0.55*	–		
LRRT	0.90	− 0.06 ns	0.21*	0.15*	–	
LRLD	0.98	− 0.07 ns	0.23*	0.15*	0.75*	–

LM1 and LM2—resistance to the coffee leaf miner; LRRT—resistance to coffee leaf rust by the reaction type and LRLD—resistance to coffee leaf rust by the lesion density. h^2: heritability; ns: not significant; * means significant at 5% by the t-test

total of 44 significant associations were identified for all traits.

Initially, twelve potential markers were significantly associated ($p\text{-value} < 6.88 \times 10^{-6}$) with resistance to biotic agents (LM1, LM2, LRRT and LRLD) (see Fig. 3 and Table 2). From this group, five SNPs are associated to resistance to coffee leaf miner (LM1 and LM2), and two are located on chromosomes 2C and 11C of the of *C. canephora* genome and three on chromosomes 2E, 9E and 11E of the *C. eugenioides* genome (Fig. 3a–b). As for the seven SNPs associated with resistance to coffee leaf rust (LRRT and LRLD), five were identified on chromosomes 1C, 2C and 6C of the *C. canephora* genome and two on chromosome 3E of the *C. eugenioides* genome (Fig. 3c–d).

The markers data made it possible to calculate the variance explained by these SNPs, allowing to estimate the inheritance of the traits in a study based on molecular markers. Analyzing the coffee leaf miner's resistance traits, we observed that the sums of the estimates of significant SNPs' heritability for each evaluation period were 0.19 (LM1) and 0.25 (LM2). In the case of traits related to coffee leaf rust, LRRT, and LRLD, the estimates' sums were 0.18 and 0.22, respectively. Moreover, most of the SNPs associated with resistance to biotic stresses had a negative allelic effect (−0.16 to −0.75) and they represent an increase in resistance to coffee leaf miner and coffee leaf rust (Table 2). Information about reference and alternative allele from significant associations for resistance to leaf miner and coffee rust are shown in Table 2.

The annotation using BLASTn searches of the twelve genomic regions that flank the SNPs associated with resistance to coffee leaf miner (LM1 and LM2) and to coffee leaf rust (LRRT and LRLD) identified four candidate genes, three of them related to plant's response to biotic stress (LOC113730621, LOC113716433 and LOC113719164) and one gene

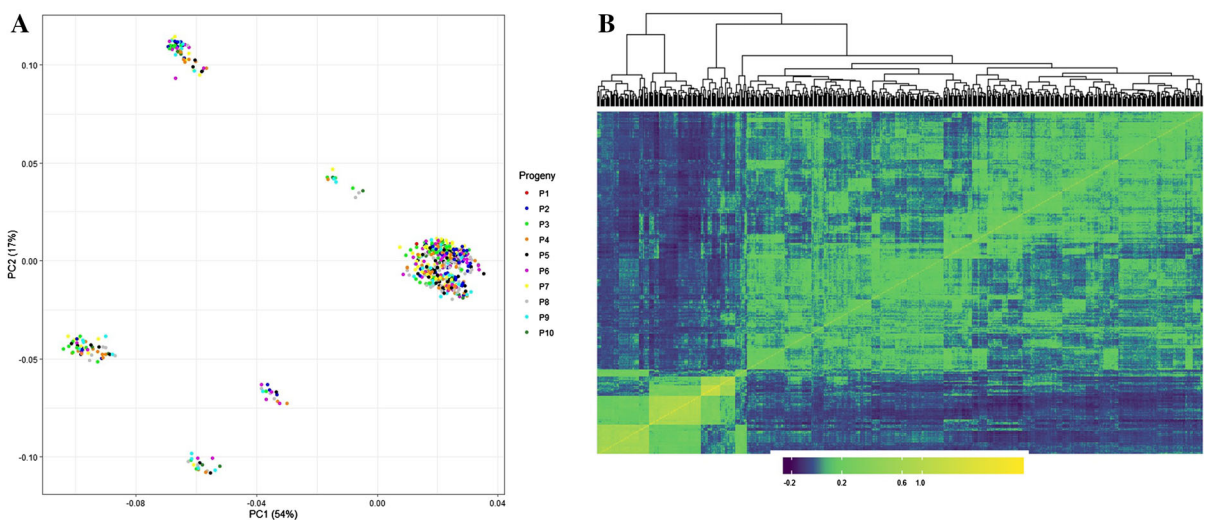
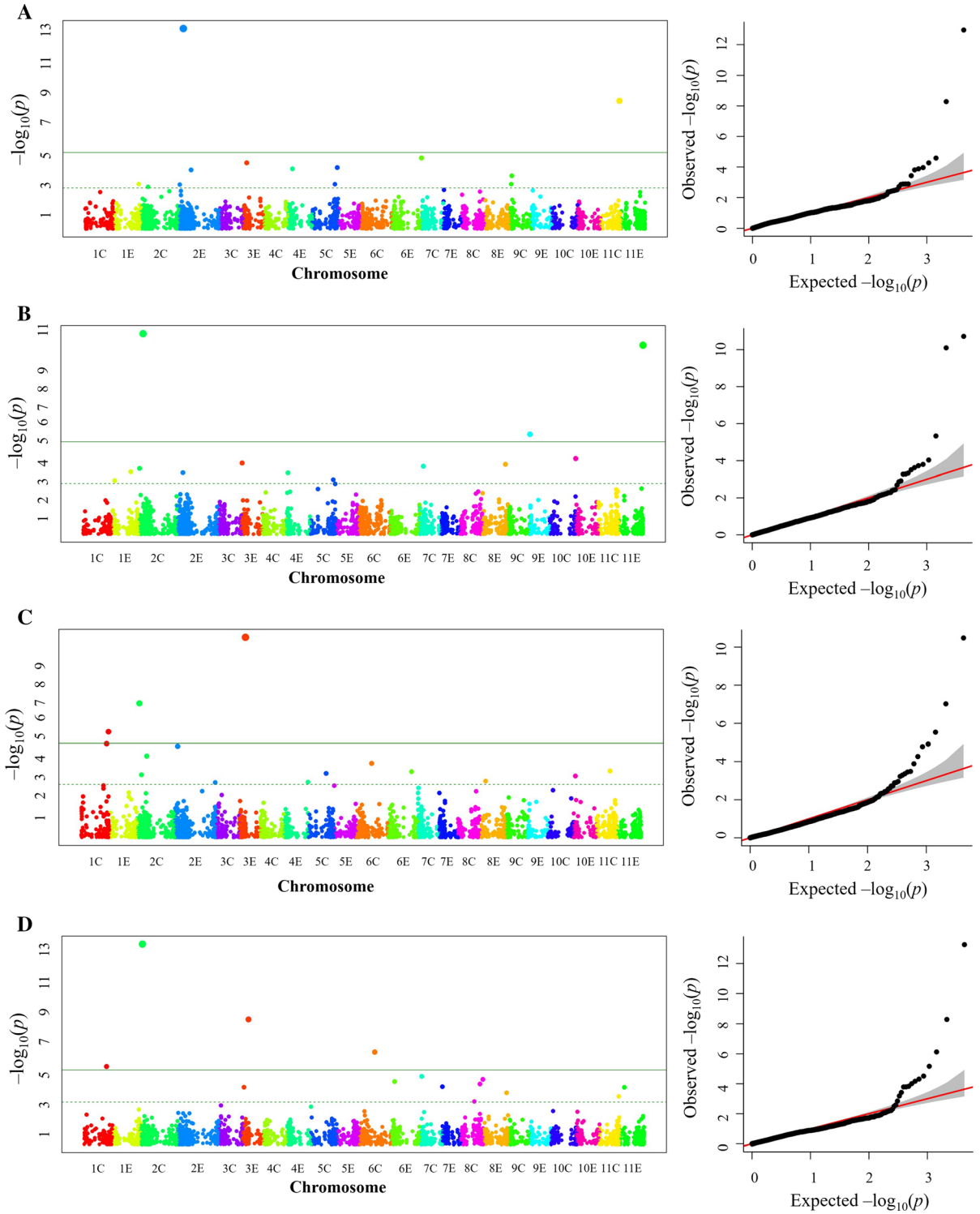


Fig. 2 Graphical representation of the population structure of ten progenies of Arabica coffee. **a** Heatmap of the kinship matrix analysis (K) and (B) Principal Component Analysis (PCA) containing two components, PC1(54%) and PC2 (17%)


A
B
C
D
−log10(p)
Chromosome
1C 1E 2C 2E 3C 3E 4C 4E 5C 5E 6C 6E 7C 7E 8C 8E 9C 9E 10C 10E 11C 11E
Observed −log10(p)
Expected −log10(p)

◀**Fig. 3** Genomic association analysis for resistance to the coffee leaf miner (LM1 and LM2), resistance to coffee leaf rust by the reaction type (LRRT) and the lesion density (LRLD). Manhattan plots for LM1, LM2, LRRT, and LRLD are shown in A, B, C, and D, respectively. QQ-plots for LM1, LM2, LRRT, and LRLD are shown in D, E, F, and G, respectively. The green line corresponds to the significance limit obtained by correcting Bonferroni (0.05). The 22 chromosomes of *Coffea arabica* are represented on the X-axis. The chromosomes of the subgenome of *C. canephora* are represented by the letter C and *C. eugenioides* by the letter E

involved in sucrose metabolism (LOC113710314). As for the LRRT and LRLD traits, five candidate genes were annotated, three genes (LOC113716485, LOC113725785 and LOC113692.424) playing regulatory roles in plant growth and development, involved in plant's responses to pathogens, respectively (Table 2); the other two genes (LOC113761697 and LOC113766162) are coincident between the two characteristics related to coffee leaf rust and resulted from the annotation of SNPs S2C_3626399 and S3E_10801345, respectively. The first, LOC113761697, is related to a protein that plays a role in meiosis, and the second, LOC113766162, refers to a protein that is involved in plant resistance to pathogens. These coincident genomic regions indicate an importance of these loci for disease defense response in coffee.

We also found 32 SNPs significantly associated (*p-value* $< 1.07 \times 10^{-5}$) with the eight scenarios designed to investigate yield (Y) and bienniality (BIEN), in which 24 mapped to the *C. eugenioides* subgenome and 8 to the *C. canephora* subgenome (Fig. 4 and Table 3). For scenarios 1, 2, 3, 4, and 5, we identified 17 significant SNPs associated with production (Fig. 4a-e). For the scenarios (6, 7 and 8), in which the MHPRVG was estimated, 15 SNPs are significantly associated with the bienniality (Fig. 4f-h). Scenario 6 presented the highest number of significant SNPs (MHPRVG for the biennium between 2005 and 2006), including six markers associated with bienniality (Fig. 4f).

The estimates of genomic heritability were low (0.01 to 0.18) for most significant SNPs, except for SNP S1E_27427517, associated with scenario 1 and exhibiting the highest value (0.44). In addition, six SNPs associated with yield (Scenario 1 to 5) showed the lowest frequency allele with a positive effect (225.91 to 836.87), enabling an increase in plants yield (g), as seen in Table 3. Information about reference and alternative allele from significative SNPs for yield and bienniality are presented in Table 3.

The annotation of the 32 SNPs associated with the different scenarios identified 25 candidate genes. These genes are related to the auxin signaling pathway, floral and seed development, abiotic stress response, plant growth and development, among other functions (Table 3).

Aiming to summarize the number of significant SNPs identified in the GWAS analysis we built a Venn Diagram using the R software (Chen and Boutros 2011). Among the 44 associated SNPs, four are coincident between at least two traits. The traits related to coffee leaf rust resistance (LR) share the same SNP (S3E_10801345) as those associated with the yield of individual years (Y2). The other three SNPs (S1E_27427517, S11E_38302756, and S8C_33803995) coincide between yield and bienniality indicator traits (Y1, Y2, and BIEN). The traits related to coffee leaf miner resistance (LM) do not share SNPs with other traits (Fig. 5).

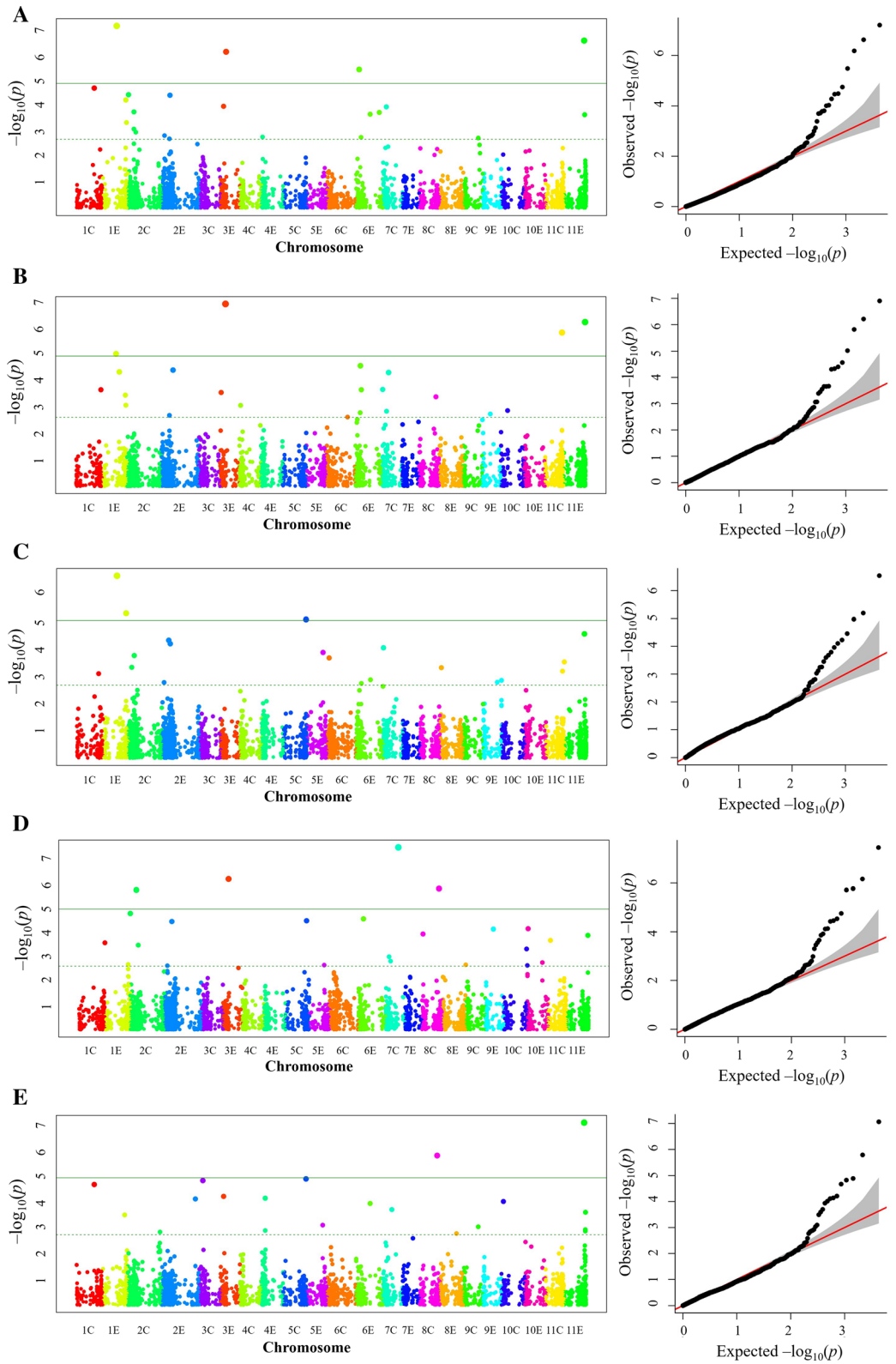
Discussion

Genomic association is a powerful approach to identify genomic regions significantly associated with phenotypic trait and has been widely applied to improve complex traits, increasing the efficiency in perennial species breeding programs. The present research stands out for being the first to identify genomic regions associated with resistance to biotic stresses caused by coffee leaf miner and coffee leaf rust, in addition to yield and bienniality in *C. arabica* progenies. The success of GWAS depends on factors such as the quality of genotypic data (SNPs). To detect high-quality SNPs, it is recommended to use a reference genome to align the sequencing reads. Such an approach minimizes one of the problems resulting from GBS sequencing: a large amount of lost data (Lipka et al. 2015). In this study, we used the *C. arabica* genome as a reference for the alignment of reads, which allowed building a matrix with a larger number of markers, and a more precise location of SNPs associated with the traits of interest.

Table 2 Annotated genes from Arabica coffee SNP markers significantly associated with the traits related to resistance to the coffee leaf miner and coffee leaf rust

Trait[1]	SNP	Chr[2]	Position	*p-value*	MAF[3]	α[4]	REF/ ALT[5]	Gene	Annotation	$h^{2(6)}$
LM1	S2E_7437676	2E	7.437.676	1.11 × 10–13	0.25	− 0.58	G/A	LOC113730621	Ubiquitin C-terminal Hydrolase (UCH)	0.13
	S11C_33085416	11C	33.085.416	5.23 × 10–9	0.46	− 0.34	G/A	LOC113716434	Transcription factor MYB124	0.06
LM2	S2C_6577134	2C	6.577.134	1.98 × 10–11	0.25	− 0.75	T/C	LOC113723788	Protein with unknown function	0.13
	S9E_641635	9E	641,635	4.66 × 10–6	0.17	− 0.24	C/T	LOC13710314	Phosphofructokinase dependent Pyrophosphate (PPi-PFP)	0.01
	S11E_42183895	11E	42.183.895	8.01 × 10–11	0.37	− 0.62	C/A	LOC113719164	L-Leucine-RichRepetition KKinase(LRR-RLK) Receiver	0.11
LRRT	S1C_48552821	1C	48.552.821	2.92 × 10–6	0.33	0.11	C/T	LOC113716485	OFP12 Transcription Repressor (Ovate Family Protein)	0.01
	S2C_3626399	2C	3.626.399	9.55 × 10–8	0.26	− 0.27	G/C	LOC113761697	DYAD Family Protein	0.07
	S3E_10801345	3E	10.801.345	3.31 × 10–11	0.32	− 0.31	C/G	LOC113766162	Allergen Pru ar 1 (PR-10 family)	0.10
LRLD	S1C_39832212	1C	39.832.212	6.88 × 10–6	0.39	− 0.16	C/T	LOC113725785	EFL1 Protein (Elongation factor like GTPase 1)	0.01
	S2C_3626399	2C	3.626.399	5.50 × 10–14	0.26	− 0.49	G/C	LOC113761697	DYAD Family Protein	0.10
	S3E_10801345	3E	10.801.345	5.16 × 10–9	0.32	− 0.45	C/G	LOC113766162	Allergen Pru ar 1 (PR-10 family)	0.09
	S6C_26336908	6C	26.336.908	7.58 × 10–7	0.06	− 0.39	A/T	LOC113692424	Metacaspase 9	0.02

[1]LM1 and LM2—resistance to the coffee leaf miner; LRRT—resistance to coffee leaf rust by the reaction type, and LRLD—resistance to leaf rust by the lesion density; [2]Chr: chromosome; [3]MAF: Minor allele frequency; α: allele substitution effect. The positive allelic effect represents an increase in susceptibility, while the negative allelic effect represents an increase in resistance to the coffee leaf miner and to coffee leaf rust; [5] REF—reference allele; ALT—alternative allele; [(6)]h^2: heritability


A
B
C
D
E
−log10(p)
Chromosome
1C 1E 2C 2E 3C 3E 4C 4E 5C 5E 6C 6E 7C 7E 8C 8E 9C 9E 10C 10E 11C 11E
Observed −log10(p)
Expected −log10(p)

◀**Fig. 4** Genomic association analysis for eight scenarios related to bienniality. **a** represents scenario 1—analysis performed for the four years together (2005, 2006, 2007, and 2008); **b**, **c**, **d** and **e** represent scenarios 2, 3, 4, and 5, respectively—each of the four years individually (2005, 2006, 2007, and 2008); **f** represents scenario 6—estimates of MHPRVG (Harmonic Mean of Relative Performance of Genetic Values) between the years of the first biennium (2005 and 2006); **g** represents scenario 7—between the years of the second biennium (2007 and 2008); and (H) represents scenario 8—which corresponds to the arithmetic mean between scenarios 6 and 7. The green line corresponds to the significance limit obtained through the Bonferroni correction (0.05). The 22 chromosomes of *Coffea arabica* are represented on the X-axis. The chromosomes of the subgenome of *C. canephora* are represented by the letter C and *C. eugenioides* by the letter E

Other relevant aspects to be considered in the GWAS analysis are population structure and genomic kinship. These parameters can affect association mapping resulting in spurious associations (Yu et al. 2006; Yang et al. 2014). In our studies these phenomena were addressed and the analyses revealed that the studied population is structured into six distinct subgroups (Fig. 2a) and these genotypes share a genomic relationship (Fig. 2b). Despite the presence of subgroups in the population, the PCA analyse did not indicate a grouping pattern, since individuals from all progenies are distributed in all groups. These facts demonstrate the need to additional analyses to a fully understanding of this population genomic structure. Probably, the inclusion of genomic data from the original parental genotypes, could clarify these

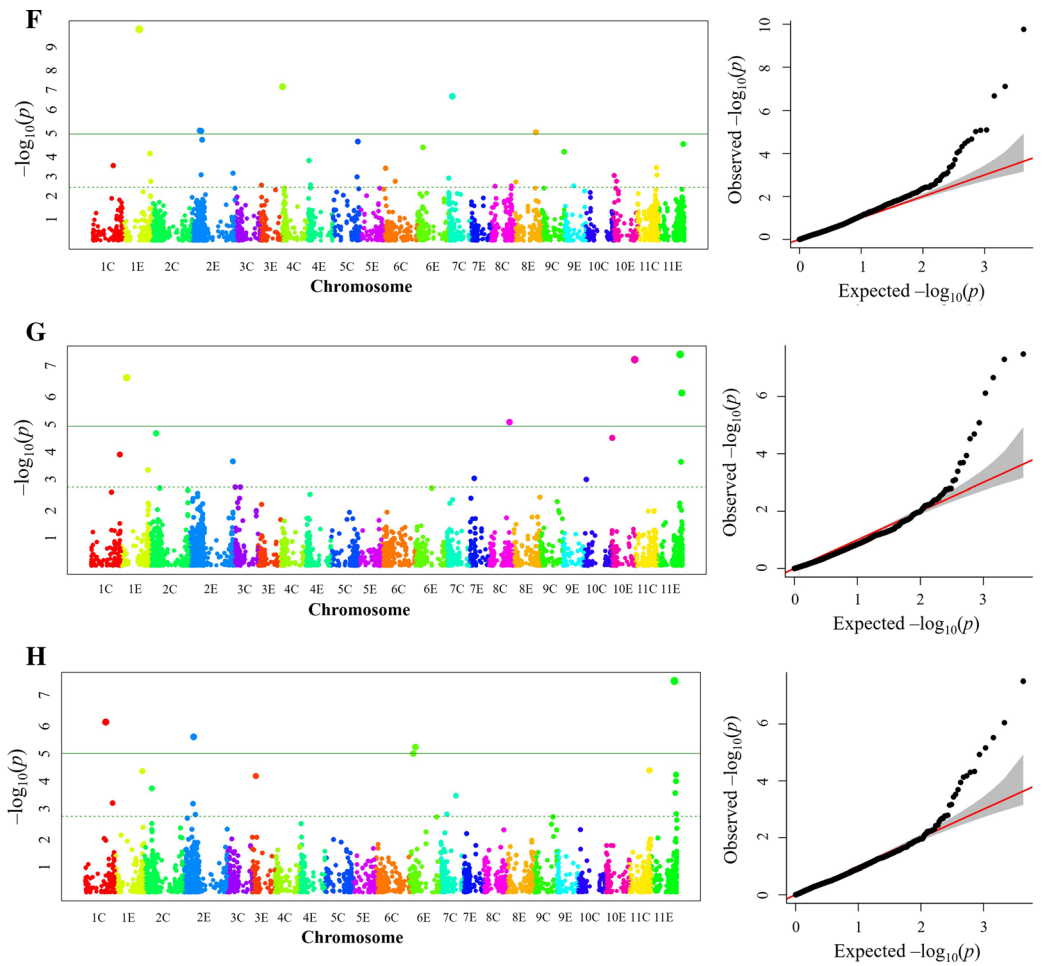


Fig. 4 continued

Table 3 Annotated genes from Arabica coffee SNPs markers significantly associated with traits with yield and bienniality of coffee progenies

Trait[1]	SNP	Chr[2]	Position	*p-value*	MAF[3]	α[4]	REF/ ALT[5]	Gene	Annotation	$h^{2(6)}$
Scenario 1	S1E_27427517	1E	27.427.517	6.26 × 10–8	0.07	− 1239.13	C/G	LOC113720800	Protein with unknown function	0.44
	S3E_10773689	3E	10.773.689	6.63 × 10–7	0.07	− 551.29	C/A	LOC1137362	PCNT115 auxin-induced protein	0.09
	S6E_6097701	6E	6.097.701	3.33 × 10–6	0.06	520.97	G/A	LOC113695394	Protein with unknown function	0.07
	S11E_38302756	11E	38.302.756	2.39 × 10–7	0.16	− 319.47	A/G	LOC113717627	Zinc-finger protein	0.06
Scenario 2	S1E_27427517	1E	27.427.517	9.60 × 10–6	0.07	− 296.42	C/G	LOC113720800	Protein with unknown function	0.03
	S3E_10801345	3E	10.801.345	1.00 × 10–7	0.33	260.34	C/G	LOC113766162	Allergen Pru ar 1 (PR-10 family)	0.07
	S11C_32570873	11C	32.570.873	1.50 × 10–6	0.36	− 173.70	A/G	LOC113748804	Transcription factor MYB16	0.03
	S11E_41239911	11E	41.239.911	6.00 × 10–7	0.29	− 250.01	G/C	LOC113753167	3-oxoacil-ACP reductase	0.06
Scenario 3	S1E_27427517	1E	27.427,517	3.00 × 10–7	0.07	− 2107.25	C/G	LOC113720800	Protein with unknown function	0.13
	S1E_44990492	1E	44.990.492	6.50 × 10–6	0.48	− 537.98	G/C	LOC113696126	ABI3 transcription factor containing domain B3	0.03
	S5C_43673515	5C	43.673.515	1.07 × 10–5	0.26	− 498.92	C/T	LOC113690486	A protein of the BPI/LBP family	0.02
Scenario 4	S2C_11931479	2C	11.931.479	1.93 × 10–6	0.46	225.91	C/T	LOC113726070	Glycogen phosphorylase 1	0.01
	S3E_10801345	3E	10.801.345	7.00 × 10–7	0.33	473.01	C/G	LOC113766162	Allergen Pru ar 1 (PR-10 family)	0.05
	S7C_24834415	7C	24.834.415	3.52 × 10–8	0.07	− 565.45	T/A	LOC113700048	Protein with unknown function	0.02
	S8C_32644919	8C	32.644.919	1.70 × 10–6	0.06	475.50	A/T	LOC113707232	BSPA subfamily protein	0.01
Scenario 5	S8C_33803995	8C	33.803.995	1.63 × 10–6	0.44	836.87	A/G	LOC113707316	EPSIN 1 Protein	0.04
	S11E_38302756	11E	38.302.756	8.60 × 10–8	0.16	− 726.62	A/G	LOC113717627	Zinc-finger protein	0.02
Scenario 6	S1E_27427517	1E	27.427.517	1.74 × 10–10	0.07	− 0.24	C/G	LOC113720800	Protein with unknown function	0.05
	S2E_12916574	2E	12.916.574	8.05 × 10–6	0.30	0.04	T/A	LOC113731095	Leucine-Rich Repetition Kinase (LRR-RLK) Receiver	0.09
	S2E_15430945	2E	15.430.945	8.40 × 10–6	0.49	− 0.24	G/A	LOC113762445	Protein with unknown function	0.18
	S4C_369260	4C	369,260	7.85 × 10–8	0.06	0.11	C/T	LOC113739559	Plasma membrane ATPase	0.02
	S7C_6467915	7C	6.467.915	2.18 × 10–7	0.49	0.31	C/A	LOC113699003	Sulfite exporter (TauE/SafE family protein)	0.18
	S8E_35577274	8E	35.577.274	9.60 × 10–6	0.09	− 0.07	T/A	LOC113706425	A protein of the PDR family	0.04
Scenario 7	S1E_9517182	1E	9.517.182	2.21 × 10–7	0.28	0.05	A/G	LOC113767523	Protein kinase serine/threonine with S-like receptor and G-type lectin	0.01
	S8C_33803995	8C	33.803.995	8.34 × 10–6	0.44	0.07	A/G	LOC113707316	EPSIN 1 Protein	0.04
	S10E_38798604	10E	38.798.604	5.09 × 10–8	0.10	− 0.10	A/G	LOC113712470	Sterol β-glucosyltransferase/ Subfamily protein UGT80B1	0.03
	S11E_38302756	11E	38.302.756	3.32 × 10–8	0.16	− 0.07	A/G	LOC113717627	Zinc-finger protein	0.01
	S11E_40942921	11E	40.942.921	7.70 × 10–7	0.12	− 0.07	C/T	LOC113718649	Protein associated with multiple drug resistance (MRP)	0.01

Table 3 continued

Trait[1]	SNP	Chr[2]	Position	*p-value*	MAF[3]	α[4]	REF/ ALT[5]	Gene	Annotation	$h^{2(6)}$
Scenario 8	S1C_34717561	1C	34.717.561	8.98×10–7	0.12	− 0.06	T/C	LOC113724164	Proteins of SSU processomo	0.02
	S2E_15190452	2E	15.190.452	3.01×10–6	0.06	0.08	G/A	LOC113731332	Protein of theTIC62 family	0.02
	S6E_9704089	6E	9.704.089	6.93×10–6	0.37	− 0.03	T/C	LOC113693770	Bifunctional protein (BIO3—BIO1)	0.00
	S11E_38302756	11E	38.302.756	3.15×10–8	0.16	− 0.06	A/G	LOC113717627	Zinc-finger protein	0.02

[1] Scenario 1—Analysis conducted for the four years together (2005. 2006. 2007, and 2008); Scenario 2, 3, 4 and 5—For each of the four years individually (2005, 2006, 2007 and 2008); Scenario 6—MHPRVG between years 2005 and 2006; Scenario 7—MHPRVG between years 2007 and 2008; and Scenario 8—Arithmetic mean between scenarios 6 and 7; [2]Chr: chromosome; [3]MAF: Minor allele frequency; [4]α: allele substitution effect. The positive allelic effect represents an increase in plant yield, while the negative allelic effect represents a reduction in plant yield; [5] REF—reference allele; ALT—alternative allele; [(6)]h^2: heritability

phenomena. Furthermore, these results point to the need to use the population structure (Q matrix) and kinship relationship (K matrix) as covariates in the GWAS models.

The Q-Q plots show the divergence between the observed significance values and those expected in the test's null hypothesis (diagonal of the graph). We verified that the distributions were uniform, that is, most of the points are located on the diagonal line with few deviations in the upper right corner of the graphs, confirming the significant associations observed in the Manhattan plots. In addition, there is a small increase in p-values in the lower corner of the graphs, showing that the models used were adequate to weigh the false positives resulting from the population structure and kinship (Yu et al. 2006).

In coffee plants, leaf miner and leaf rust are considered the most important phytosanitary problems, and so far, there is no cultivar resistant to both diseases. Therefore, identifying molecular markers that are related to plant's defense mechanisms against pests and diseases can help in the development of resistant cultivars (DaMatta et al. 2007; Lashermes et al. 2009). These defense mechanisms may involve: activation of defense genes, phytohormones signaling (auxin, jasmonic acid, salicylic acid, and others), PAMPs (Pathogen-Associated Molecular Patterns) and HAMPs (Herbivore-Associated Molecular Patterns) recognition, among others. Probably the genomic regions associated with resistance to coffee leaf miner and coffee leaf rust identified here may trigger a defense response in coffee plants.

Following the pipeline, the search for candidate genes from SNPs associated with resistance to coffee leaf miner (LM1 and LM2) resulted in four genes with known functions. Two genes (LOC113730621 and LOC113716433) were noted in the ubiquitin carboxy-terminal protein hydrolase and the transcription factor MYB124, respectively. These two proteins are related to jasmonic acid (JA) metabolism (Campos et al. 2018; Shen et al. 2018). In plants, the JA signaling pathway plays an important role in the response and resistance to biotic agents (Howe et al. 2018). Another candidate gene, LOC113719164, corresponds to the LRR-RLK (Leucine-Rich Repeat receptor-like protein kinase), belonging to the receptor-like kinase (RLKs) protein family, with leucine-rich repetitions (LRRs). This class of proteins (LRR-RLKs) has pattern recognition receptors (PRRs) that recognize

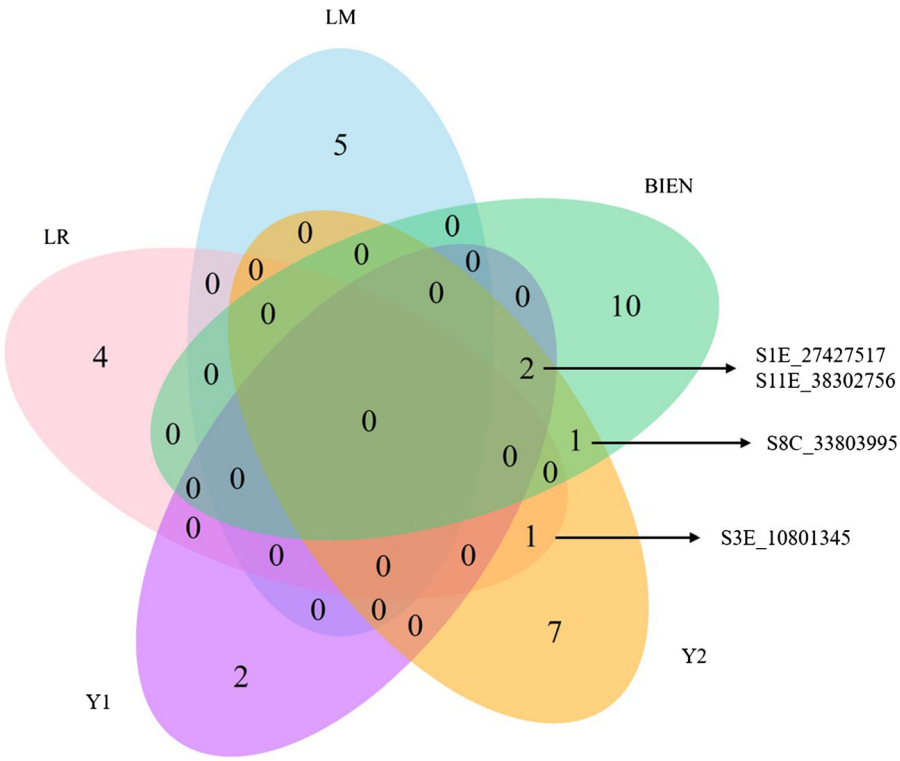


Fig. 5 Markers identified as significant coincidences between the analyzed traits. LM: traits related to resistance to coffee leaf miner (LM1 and LM2); FE: traits related to resistance to coffee leaf rust (LRRT and LRLD); YIELD1: scenario 1; YIELD2: scenarios 2 to 5; BIEN: scenarios 6 to 8

pathogen-associated molecular patterns (PAMPs) and herbivore-associated molecular patterns (HAMPs). In addition, several RLK family genes are part of resistance mechanism to pathogens (Sekhwal et al. 2015) and insects (Liu et al. 2015). As shown in Table 2, there are no coincident genomic regions between the characteristics LM1 and LM2, probably because the evaluations for resistance to the insect were performed in different phenological phases: fruiting (LM1); flowering and leaf development (LM2).

We discovered five candidate genes close to significant SNPs regarding the LRRT and LRLD traits linked to coffee leaf rust resistance. The candidate gene LOC113716485 corresponds to the transcription repressor OFP12. The OVATE family protein (OFP) is composed of plant-specific transcriptional regulators with a well-known role in several processes of plant development, growth and resistance (Huang et al. 2013; Ding et al. 2020). This protein family operates in signaling pathways of phytohormones, such as auxin, that can works as a negative regulator of resistance to biotrophic pathogens, such as coffee leaf rust (Wang et al. 2007; Mutka et al. 2013). In addition, two candidate genes (LOC113761697 and LOC113766162) were coincident between the LRRT and LRLD traits, possibly due to the high correlation observed between them (Table 1). The first gene encodes a DYAD family protein, which plays a role in chromosome pairing (Li et al. 2005). In this sense, recent study demonstrate that the SWITCH 1/DYAD gene acts during meiosis in *Arabidopsis thaliana* (Yang et al. 2019). The second candidate gene (LOC113766162), *Pru ar 1*, belongs to the PR-10 protein family. The PR genes degrade cell walls of pathogens, inhibiting their growth and playing a key role in the plant's defense pathways (Chen et al. 2010; Fan et al. 2015; Zhang et al. 2017). These coincident genomic regions between LRRT and LRLD indicate the importance of these loci for coffee leaf rust resistance.

In general, yield is a complex trait, influenced by many genes. Carvalho et al. (2020), for example, used the same breeding population of *C. arabica* analyzed here and identify a low heritability estimate for this trait, evidencing the referred complexity. Although we identified a large number of candidate genes related to production and biennial scenarios, we will discuss here those relevant to the traits in question.

We identified three genes related to proteins that have a function in phytohormones signaling pathways. The gene LOC1137362 (scenario 1) encodes an auxin-induced protein (PCNT115) involved in the signaling pathway of this plant hormone (Gao et al. 2014). The auxin slows down the senescence of fruits, besides playing a fundamental role in their growth and development (Pramanik and P. Mohapatra 2017).

The other two genes are related to proteins from the abscisic acid (ABA) signaling pathway. Some studies suggest that an interaction between ABA and sucrose may occur during rice grain filling phase (Chen et al. 2019; Rezaul et al. 2019). The gene LOC113696126 (scenario 3) has homology to a transcription factor insensitive to abscisic acid (ABI3). Radchuk et al. (2010) reported that the ABI3 gene was expressed late in pea seeds delaying maturation. The gene (LOC113767523—scenario 7) finally encodes a serine/threonine kinase protein with an S-like receptor and G-type lectin. According to Sun et al. (2013), the homologous gene in *Glycine soybean* (*GsSRK*) induces ABA's biosynthesis.

Besides those genes mentioned above, we highlight the gene LOC113717627, which coincides in some scenarios (1, 5, 7, and 8) referring to yield and bienniality. This gene belongs to a family of Zinc finger proteins involved in several essential biological functions, including floral development (Dinneny 2004; Ohno 2004; Xiao et al. 2009) and abiotic stress response (Zhang et al. 2016). These coincident candidate genes may indicate a stability and importance of these genomic regions for yield and bienniality. Another important gene, LOC113766162, was also simultaneously associated with two scenarios (2 and 4) related to production and with LRRT and LRLD traits. This gene encodes the protein *Pru ar 1* involved in resistance to biotic stresses (Fan et al. 2015). Some authors indicate that coffee leaf rust is related to yield, i.e., susceptible cultivars, the incidence and severity of the disease increases in years of high yield (Zambolim 2016; Toniutti et al. 2017).

The significant SNP markers identified in this work were annotated and may play an important role in plants' response to biotic and abiotic stresses. Therefore, when validated, these markers may constitute target regions to assist conventional coffee breeding programs by applying molecular improvement techniques and biotechnology. The SNPs identified here can increase the allele frequency related to pest and disease resistance and yield gains in breeding populations by introducing genotypes that possess these alleles and thus target crosses more efficiently. In addition, the use of these markers may potentially identify the presence of favorable/unfavorable alleles in segregating populations under selection. In this way, a large number of resistant and productive individuals can be early selected, without the need for intense field evaluations over several years, making the development of new coffee cultivars faster and more efficient.

Several GWAS studies on cultivated plant species have highlighted the possibility of applying this approach in traditional breeding programs via SAM. However, the use of GWAS as a routine in these programs faces some difficulties, as it requires a fast and low-cost genotyping methodology (Pantalião et al. 2016). Currently, the traditional breeding of coffee has well-established procedures to obtain cultivars. However, the process is still awfully long. From this perspective, genomic tools can reduce the selection cycles necessary for launching productive coffee cultivars resistant to the main biotic agents. Recently, the advancement of genomics has made possible the sequencing of the complete genome of the most commercially important species of the genus *Coffea* (*C. canephora and C. arabica*) and allowed the identification of SNPs related to several agronomic traits.

Conclusion

The association mapping carried out on a population of *C. arabica* allowed the identification of 44 SNPs significantly associated with resistance to biotic stresses, yield, and bienniality. Of these, three were coincident between at least two of the evaluated traits. The genomic region LOC113766162 stands out for being shared between yield and the traits related to coffee leaf rust resistance. Several of these SNPs have

been noted in genomic regions involved with biological processes important for coffee breeding. Therefore, those SNPs can help breeding programs to develop strategies that accelerate the process through marker-assisted selection.

Acknowledgements The authors are grateful to the National Council for Scientific and Technological Development (CNPq) for research fellowship (OGF CNPq DT 307.610/2020-9), Coordination of Superior Level Staff Improvement (CAPES), Foundation for Research Support of the State of Minas Gerais (FAPEMIG), Instituto Nacional de Ciência e Tecnologia do Café (INCT – Café), Brazilian Agricultural Research Corporation (EMBRAPA-Coffee), Secretaria de Agricultura e Abastecimento do Estado de São Paulo (SAASP), and Coffee Research Consortium for the financial support.

References

Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next-generation reference panels. Am J Hum Genet 103:338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Campos ML, De Souza CM, De Oliveira KBS et al (2018) The role of antimicrobial peptides in plant immunity. J Exp Bot 69:4997–5011. https://doi.org/10.1093/jxb/ery294

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

Ceccarelli S (2015) Efficiency of plant breeding. Crop Sci 55:87–97. https://doi.org/10.2135/cropsci2014.02.0158

Chen H, Boutros PC (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinform. https://doi.org/10.1186/1471-2105-12-35

Chen T, Li G, Islam MR et al (2019) Abscisic acid synergizes with sucrose to enhance grain yield and quality of rice by improving the source-sink relationship. BMC Plant Biol 19:1–17. https://doi.org/10.1186/s12870-019-2126-y

Chen Z-Y, Brown RL, Damman KE, Cleveland TE (2010) PR10 expression in maize and its effect on host resistance against Aspergillus flavus infection and aflatoxin production. Mol Plant Pathol. https://doi.org/10.1111/j.1364-3703.2009.00574.x

DaMatta FM, Ronchi CP, Maestri M, Barros RS (2007) Ecophysiology of coffee growth and production. Braz J Plant Physiol. https://doi.org/10.1590/S1677-04202007000400014

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

Dantas J, Motta I, Vidal L, et al (2020) A Comprehensive Review of the Coffee Leaf Miner Leucoptera Coffeella (Lepidoptera: Lyonetiidae), With Special Regard to Neotropical Impacts, Pest Management and Control. Preprints

Ding B, Hu C, Feng X et al (2020) Systematic analysis of the OFP genes in six Rosaceae genomes and their roles in stress response in Chinese pear (Pyrus bretschneideri). Physiol Mol Biol Plants. https://doi.org/10.1007/s12298-020-00866-3

Dinneny JR (2004) The role of JAGGED in shaping lateral organs. Development. https://doi.org/10.1242/dev.00949

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus (madison) 12:13–15

Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:1–10. https://doi.org/10.1371/journal.pone.0019379

Eskes A, Braghini M (1981) Assessment methods for resistance to coffee leaf rust (Hemileia vastatrix Berk. and Br.). FAO plant Prot Bull

Fan S, Jiang L, Wu J et al (2015) A novel pathogenesis-related class 10 protein Gly m 4l, increases resistance upon phytophthora sojae infection in soybean (Glycine max [L.] Merr.). PLoS One. https://doi.org/10.1371/journal.pone.0140364

Gao C, Wang Y, Shen Y et al (2014) Oil accumulation mechanisms of the oleaginous microalga *Chlorella prototheco*ides revealed through its genome, transcriptomes, and proteomes. BMC Genom 15:1–14. https://doi.org/10.1186/1471-2164-15-582

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. 1–9

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

Guerreiro Filho O, Silvarolla MB, Eskes AB (1999) Expression and mode of inheritance of resistance in coffee to leaf miner *Perileucoptera coffeella*. Euphytica 105:7–15. https://doi.org/10.1023/A:1003427613071

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol. https://doi.org/10.1007/s11103-005-0257-z

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models the MCMCglmm R package. J Stat Softw 33:1–22. https://doi.org/10.18637/jss.v033.i02

Howe GA, Major IT, Koo AJ (2018) Modularity in jasmonate signaling for multistress resilience. Annu Rev Plant Biol 69:387–415. https://doi.org/10.1146/annurev-arplant-042817-040047

Huang Z, Van Houten J, Gonzalez G et al (2013) Genome-wide identification, phylogeny and expression analysis of SUN, OFP and YABBY gene family in tomato. Mol Genet Genom. https://doi.org/10.1007/s00438-013-0733-0

ICO (2020) International Coffee Organization (ICO), Monthly Coffee Market Report (2019/20). http ://www.ico.org/Market-Report-19-20-e.asp

Imai A, Nonaka K, Kuniga T et al (2018) Genome-wide association mapping of fruit-quality traits using genotyping-by-sequencing approach in citrus landraces, modern cultivars, and breeding lines in Japan. Tree Genet Genom. https://doi.org/10.1007/s11295-018-1238-0

Kainer D, Padovan A, Degenhardt J et al (2019) High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*. New Phytol 223:1489–1504. https://doi.org/10.1111/nph.15887

Kolkman JM, Berry ST, Leon AJ et al (2007) Single nucleotide polymorphisms and linkage disequilibrium in sunflower. Genetics. https://doi.org/10.1534/genetics.107.074054

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

Lashermes P, Bertrand B, Etienne H (2009) Breeding coffee (Coffea arabica) for sustainable production In breeding plantation tree crops: tropical species. Springer, New York

Lashermes P, Combes MC, Hueber Y et al (2014) Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. Plant J 78:674–685. https://doi.org/10.1111/tpj.12505

Li W, Yang X, Lin Z et al (2005) The *AtRAD51C* gene is required for normal meiotic chromosome synapsis and double-stranded break repair in Arabidopsis. Plant Physiol 138:965–976. https://doi.org/10.1104/pp.104.058347

Lipka AE, Kandianis CB, Hudson ME et al (2015) From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol 24:110–118. https://doi.org/10.1016/j.pbi.2015.02.010

Liu X, Huang M, Fan B et al (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLOS Genet 12:e1005767. https://doi.org/10.1371/journal.pgen.1005767

Liu Y, Wu H, Chen H et al (2015) A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. Nat Biotechnol 33:301–307. https://doi.org/10.1038/nbt.3069

Minamikawa MF, Nonaka K, Kaminuma E et al (2017) Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. Sci Rep 7:1–2. https://doi.org/10.1038/s41598-017-05100-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 Genom 12:1–17. https://doi.org/10.1007/s11295-015-0927-1

Mutka AM, Fawley S, Tsao T, Kunkel BN (2013) Auxin promotes susceptibility to *Pseudomonas syringae* via a mechanism independent of suppression of salicylic acid-mediated defenses. Plant J 74:746–754. https://doi.org/10.1111/tpj.12157

Nguyen KL, Grondin A, Courtois B, Gantet P (2019) Next-generation sequencing accelerates crop gene discovery. Trends Plant Sci 24:263–274. https://doi.org/10.1016/j.tplants.2018.11.008

Ohno CK (2004) The Arabidopsis JAGGED gene encodes a zinc finger protein that promotes leaf tissue development. Development. https://doi.org/10.1242/dev.00991

Pantalião GF, Narciso M, Guimarães C et al (2016) Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. Genetica 144:651–664. https://doi.org/10.1007/s10709-016-9932-z

Pereira DR, Nadaleti DH, Rodrigues EC et al (2021) Genetic and chemical control of coffee rust (Hemileia vastatrix) impacts on coffee (Coffea arabica L) quality. J Sci Food Agric. https://doi.org/10.1002/jsfa.10914

Pestana KN, Capucho AS, Caixeta ET et al (2015) Inheritance study and linkage mapping of resistance loci to *Hemileia vastatrix* in Híbrido de Timor UFV 443–03. Tree Genet Genom 11:1–13. https://doi.org/10.1007/s11295-015-0903-9

Pramanik K, Mohapatra P (2017) Role of auxin on growth, yield and quality of tomato - A review. Int J Curr Microbiol Appl Sci 6:1624–1636. https://doi.org/10.20546/ijcmas.2017.611.195

Radchuk R, Conrad U, Saalbach I et al (2010) Abscisic acid deficiency of developing pea embryos achieved by immunomodulation attenuates developmental phase transition and storage metabolism. Plant J 64:715–730. https://doi.org/10.1111/j.1365-313X.2010.04376.x

Rena A, Maestri M (1987) Ecofisiologia do cafeeiro. In: Ecofisiologia da produção agricola. Piracicaba, pp 249-undefined

Resende MDV (2002) Genética biométrica e estatística no melhoramento de plantas perenes. Embrapa Informação Tecnológica, Brasília

Rezaul IM, Baohua F, Tingting C et al (2019) Abscisic acid prevents pollen abortion under high-temperature stress by mediating sugar metabolism in rice spikelets. Physiol Plant 165:644–663. https://doi.org/10.1111/ppl.12759

Richards JK, Friesen TL, Brueggeman RS (2017) Association mapping utilizing diverse barley lines reveals net form net blotch seedling resistance/susceptibility loci. Theor Appl Genet 130:915–927. https://doi.org/10.1007/s00122-017-2860-1

Romero Navarro JA, Wilbert P-M, Arciniegas-Leal A et al (2017) Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. Front Plant Sci. https://doi.org/10.3389/fpls.2017.01905

Santana GC, Pereira LFP, 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

Sekhwal MK, Li P, Lam I et al (2015) Disease resistance gene analogs (RGAs) in plants. Int J Mol Sci 16:19248–19290. https://doi.org/10.3390/ijms160819248

Setotaw TA, Caixeta ET, Pereira AA et al (2013) Coefficient of parentage in *Coffea arabica L cultivars grown in Brazil*. Crop Sci. https://doi.org/10.2135/cropsci2012.09.0541

Shen X-J, Wang Y-Y, Zhang Y-X et al (2018) Overexpression of the wild soybean R2R3-MYB transcription factor GsMYB15 enhances resistance to salt stress and helicoverpa armigera in transgenic arabidopsis. Int J Mol Sci. https://doi.org/10.3390/ijms19123958

Singh VK, Sharma V, Paswan SK et al (2016) Study on genetic variability, heritability and genetic advance for yield and its contributing traits in linseed (Linum usitatissimum L). Curr Adv Agric Sci Int J. https://doi.org/10.5958/2394-4471.2016.00046.0

Sousa TV, Caixeta ET, Alkimim ER 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 Genom. 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 9:1–12. https://doi.org/10.3389/fpls.2018.01934

Sun X-L, Yu Q-Y, Tang L-L et al (2013) GsSRK, a G-type lectin S-receptor-like serine/threonine protein kinase, is a positive regulator of plant tolerance to salt stress. J Plant Physiol. https://doi.org/10.1016/j.jplph.2012.11.017

Tibbs Cortes L, Zhang Z, Yu J (2021) Status and prospects of genome-wide association studies in plants. Plant Genome 14:1–17. https://doi.org/10.1002/tpg2.20077

Toniutti L, Breitler J-C, Etienne H et al (2017) Influence of environmental conditions and genetic background of arabica coffee (C arabica L) on leaf rust (Hemileia vastatrix) pathogenesis. Front Plant Sci. https://doi.org/10.3389/fpls.2017.02025

Tran HTM, Furtado A, Vargas CAC et al (2018) SNP in the Coffea arabica genome associated with coffee quality. Tree Genet Genomes. https://doi.org/10.1007/s11295-018-1282-9

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

Volsi B, Telles TS, Caldarelli CE, da Camara MRG (2019) The dynamics of coffee production in Brazil. PLoS One 14:e0219742

Vos PG, Paulo MJ, Voorrips RE et al (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theor Appl Genet. https://doi.org/10.1007/s00122-016-2798-8

Wang S, Chang Y, Guo J, Chen JG (2007) Arabidopsis ovate family protein 1 is a transcriptional repressor that suppresses cell elongation. Plant J 50:858–872. https://doi.org/10.1111/j.1365-313X.2007.03096.x

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York

Wimmer V, Albrecht T, Auinger H-J, Schön C-C (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts335

Xiao H, Tang J, Li Y et al (2009) *STAMENLESS 1*, encoding a single C2H2 zinc finger protein, regulates floral organ identity in rice. Plant J 59:789–801. https://doi.org/10.1111/j.1365-313X.2009.03913.x

Yang J, Lee SH, Goddard ME, Visscher PM (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol Biol 1019:215–236. https://doi.org/10.1007/978-1-62703-447-0_9

Yang C, Hamamura Y, Sofroni K et al (2019) SWITCH 1/DYAD novel WINGS APART-LIKE antagonist that maintains sister chromatid cohesion in meiosis. bioRxiv. https://doi.org/10.1101/551317

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

Zambolim L (2016) Current status and management of coffee leaf rust in Brazil. Trop Plant Pathol 41:1–8. https://doi.org/10.1007/s40858-016-0065-9

Zhang D, Tong J, Xu Z et al (2016) Soybean C2H2-type zinc finger protein GmZFP3 with conserved QALGGH motif negatively regulates drought responses in transgenic *Arabidopsis*. Front Plant Sci 7:1–9. https://doi.org/10.3389/fpls.2016.00325

Zhang H, Chen L, Sun Y et al (2017) Investigating proteome and transcriptome defense response of apples induced by *Yarrowia lipolytica*. Mol Plant-Microbe Interact 30:301–311. https://doi.org/10.1094/MPMI-09-16-0189-R

Zhang Y, Liu P, Zhang X et al (2018) Multi-locus genome-wide association study reveals the genetic architecture of stalk lodging resistance-related traits in maize. Front Plant Sci 9:611. https://doi.org/10.3389/fpls.2018.00611