BMC Plant Biology

RESEARCH

Open Access

Genome-wide association insights into the genomic regions controlling vegetative and oil production traits in *Acrocomia aculeata*

Evellyn G. O. Couto[1*], Jonathan A. Morales-Marroquín[1], Alessandro Alves-Pereira[2], Samuel B. Fernandes[3], Carlos Augusto Colombo[4], Joaquim Adelino de Azevedo-Filho[4], Cassia Regina Limonta Carvalho[4] and Maria Imaculada Zucchi[1,5*]

Abstract

Background Macauba (*Acrocomia aculeata*) is a non-domesticated neotropical palm that has been attracting attention for economic use due to its great potential for oil production comparable to the commercially used oil palm (*Elaeis guineensis*). The discovery of associations between quantitative trait loci and economically important traits represents an advance toward understanding its genetic architecture and can contribute to accelerating macauba domestication. Pursuing this advance, this study performs single-trait and multi-trait GWAS models to identify candidate genes associated with vegetative and oil production traits in macauba. Eighteen phenotypic traits were evaluated from 201 palms within a native population. Genotyping was performed with SNP markers, following the protocol of genotyping-by-sequencing. Given that macauba lacks a reference genome, SNP calling was performed using three different strategies: using i) de novo sequencing, ii) the *Elaeis guineenses* Jacq. reference genome and iii) the macauba transcriptome sequences. After quality control, we identified a total of 27,410 SNPs in 153 individuals for the de novo genotypic dataset, 10,444 SNPs in 158 individuals using the oil palm genotypic dataset, and 4,329 SNPs in 167 individuals using the transcriptome genotypic dataset. The GWAS analysis was then performed on these three genotypic datasets.

Results Statistical phenotypic analyses revealed significant differences across all studied traits, with heritability values ranging from 63 to 95%. This indicates that the population contains promising genotypes for selection and the initiation of breeding programs. Genetic correlations between the 18 traits ranged from -0.47 to 0.99. The total number of significant SNPs in the single-trait and multi-trait GWAS was 92 and 6 using the de novo genotypic dataset, 19 and 11 using the oil palm genotypic dataset, and 1 and 2 using the transcriptome genotypic dataset, respectively. Gene annotation identified 12 candidate genes in the single-trait GWAS and four in the multi-trait GWAS, across the 18 phenotypic traits studied, in the three genotypic datasets. Gene mapping of the macauba candidate

*Correspondence:
Evellyn G. O. Couto
evellyncoutoo@gmail.com
Maria Imaculada Zucchi
mizucchi@sp.gov.br
Full list of author information is available at the end of the article

genes revealed similarities with *Elaeis guineensis* and *Phoenix dactylifera*. The candidate genes detected are responsible for metal ion binding and transport, protein transportation, DNA repair, and other cell regulation biological processes.

Conclusions We provide new insights into genomic regions that map candidate genes associated with vegetative and oil production traits in macauba. These potential candidate genes require confirmation through targeted functional analyses in the future, and multi-trait associations need to be scrutinized to investigate the presence of pleiotropic or linked genes. Markers linked to traits of interest could serve as valuable resources for the development of marker-assisted selection in macauba for its domestication and pre-breeding.

Keywords Neotropical oil palm, Macauba, Associative mapping, Multi-trait associations, Oil content, Biofuel

Background

Macauba (*Acrocomia aculeata*) is a neotropical palm distributed from Mexico to the Northern region of Argentina [1–3]. It has significant potential for commercial exploitation due to the high oleic concentration in its fruits for energy production, food industries, pharmaceutical, and cosmetic industries [4, 5]. As a perennial plant, macauba offers long-term regular production and is well-adapted to growing in dry biomes such as bare ground, rangelands, and anthropogenic landscapes, unlike other energy crops that depend on humid ecosystems like tropical rainforests. Compared to other oilseeds, the estimated oil productivity of Brazilian populations at a density of 400 trees per hectare per year is approximately 2.5 tons. This yield is similar to oil palm (*Elaeis guineensis*) and higher than soy, cotton, and sunflower [6]. Moreover, it can be used in ecosystems programs and agroforestry systems [7]. These features make macauba a sustainable source of vegetable oil. However, the economic viability of *Acrocomia* production depends on genetic research, technological advances, and political support to develop its value chain [8]. In industry, all parts of the macauba fruit (husk, pulp, endocarp, and kernel) can be used as raw materials. Oil is extracted from both the pulp and kernel. The pulp contains up to 70% of oleic acid-rich oil [9], followed by palmitic and linoleic acid, wich are targeted by the biofuel industry. The kernel contains up to 50% oil rich in lauric acid, used in cosmetics and other saponification products [10]. The fruits are also rich in amino acids, carbohydrates, vitamins, fibers and minerals [11], making them suitable for both the animal and human food industries. Additionally, the endocarp has high lignin content, making it suitable for coal production [12]. Furthermore, macauba's nutritional profile includes antioxidants like flavonoids and phenolic acids, which may offer health benefits such as minimizing oxidative stress and lowering the risk of chronic diseases, making it valuable for the pharmaceutical industry [13, 14]. However, despite its commercial potential, macauba is considered incipiently domesticated [15, 16], and most of the oil consumed is obtained through extractivism [17].

Studies of genetic diversity can guide the selection of contrasting parents for artificial crosses, maximizing genetic gains and enhancing the efficiency of new cultivar in development [18, 19]. In this context, given that macauba is a native palm, genetic improvement is essential for developing productive varieties with commercially stable phenotypic traits. Recent studies on genetic diversity using SNP markers have evaluated *Acrocomia* spp. palms across Brazil, indicating that productive germplasm can be selected to initiate a pre-breeding cycle through targeted crosses between genetically divergent species [20, 21]. In plant breeding, the selection of parental plants with potential productivity within a natural population is based on their phenotypic traits, being that correlation estimates between these specific traits facilitate selection and enhance understanding of their genetic basis. Previous studies on macauba have observed positive genotypic correlations between vegetative and oil production traits, which can facilitate the breeding process, as breeders can employ indirect selection by focusing on a few easily measured traits [9, 22, 23]. Moreover, positive genetic correlation at the genetic study level of these traits suggest the presence of pleiotropic and/or linked genes controlling them, enabling a better understanding of the species' genetic architecture [24].

For genetic research in crop breeding and domestication studies, a high-quality reference genome at the chromosomal level is essential, but *Acrocomia* does not currently have one. However, population genomics approaches like GBS (genotyping-by-sequencing) or RAD-Seq can be employed to explore the evolutionary processes driving diversity in non-model plant species [25, 26]. Domestication often results in significant genomic changes driven by factors such as genetic drift and artificial selection. Genomic techniques are frequently used to detect selection signatures by identifying polymorphisms in interesting trait loci hidden by neutral variation. This approach has already been applied to *Acrocomia* species finding putative selection signatures in genes associated to fatty acid and carotenoid biosynthesis, as well as pathogen resistance and drought tolerance [26]. Other methodologies that can guide pre-breeding in

macauba include genetic architecture studies and association mapping, with genome-wide association studies (GWAS) being particularly notable in this case. These approaches provide genetic insights into the species and allow for the future implementation of genomic prediction technologies. In oil palm, for example, multiple genetic maps have been constructed, demonstrating that traits involved in oil production are quantitative. Molecular-assisted breeding programs have even facilitated the domestication of oil palm through fine mapping, GWAS, and cross-validation of QTLs linked to fatty acid composition and yield, [27, 28], oil palm hybrids [29], and disease resistance [30]. For example, one of the quantitative traits in oil palm that directly contributes to increased oil yield is fruit bunch weight [31]. Babu et al. [27, 28] identified numerous quantitative trait loci (QTL) associated with bunch, yield, and oil yield-related traits in oil palm through GWAS. Identifying significant loci for important economic traits is crucial for enhancing yield and oil-related traits, as well as for their use in marker-assisted selection. These studies have facilitated advances in the development of improved oil palm populations [27, 28, 31, 32], and it is such advances that we aim to achieve in the future with the present study on macauba.

In genetic architecture studies and association mapping through GWAS, molecular markers are used to map QTLs for various traits of interest, allowing a better understanding of the genetic expression of these traits [24, 32, 33]. GWAS attempts to evaluate causal relationships between genetic variants and the phenotype by surveying a genome-wide set of genetic polymorphisms in a large number of individuals [34]. Variants are normally detected using short reads and mapped to a reference genome, however common subsequences can also be directly compared between samples. Such an approach appears to be most effective when there is no reference genome assembly like macauba [35]. In the GWAS, single and multi-trait approaches are currently employed to detect cross-phenotype associations [36–38]. Some of the first GWAS models, such as general linear models and mixed-linear models (MLMs), were single-locus and single-trait, created to implement covariates along with kinship matrices [39]. These simple models resulted in false negatives caused by weakened associations due to population structure. To evaluate large datasets while also reducing false positives and negatives, multi-locus GWAS, such as FarmCPU [40] and BLINK [41], were developed. Unlike single-trait models, multi-trait models enable the quantification of simultaneous contributions of loci to multiple traits in GWAS studies, serving as an effective tool to guide research on linked or pleiotropic genes in association mapping [36, 42]. Many software packages have implemented multi-trait GWAS models [36, 43–45], and these models can often outperform univariate multi-locus models, especially when analyzing traits with low heritability [46].

The application of GWAS in macauba, particularly in its wild relatives, can enable the exploration of adaptive traits that have evolved under specific environmental pressures. Despite the challenges of studying natural populations, GWAS can uncover genetic loci that contribute to yield and oil productivity, environmental stress tolerance, and growth efficiency in poor soils. These findings can expand our understanding of the south american genetic pool for macauba cultivation, especially in regions vulnerable to climate change. Moreover, the identification of loci influencing multiple traits using multi-trait GWAS models could accelerate the development of cultivars, eventually advancing macauba from an incipiently domesticated species to a commercially viable crop with high agronomic potential. To the best of our knowledge, this study presents the first GWAS targeting genomic regions controlling vegetative and oil production traits in macauba. Results related to the genomic association of traits responsible for vegetative and oil production are of interest for understanding their genetic architecture, which would facilitate the specie's domestication and pre-breeding, similar to oil palm [27, 28]. In addition, markers detected in single-trait and multi-trait GWAS can guide studies on the detection of linked or pleiotropic genes, enriching the genetic architecture information in macauba.

Methods

Plant material

We studied a panel of 201 macauba palms from a natural population located in Dourado, São Paulo state, Brazil (geographical coordinates 22° 6′ 13" S, 48° 18′ 50" W). Palms in this population emerged through natural dispersal processes, and individuals were randomly selected, meaning that there is no specific experimental desing. The population is situated in a rural area within the Cerrado region (Fig. 1). Phenotype data were collected in October 2019 and January 2021 to represent two different production seasons (years 1 [2019/2020] and 2 [2020/2021]), and we select the palms prioritizing those with ripe fruits in October 2019.

Phenotype data

The phenotype data were divided into two categories to facilitate understanding: vegetative and oil production traits. Fig. 2A represents one of the macauba trees used in the analyses. The vegetative traits studied were: height (H in meters), stipe (trunk) diameter at breast height (DHB in centimeters), number of leaves (LN), leaf length (LL in centimeters), number of leaf needles (NN),

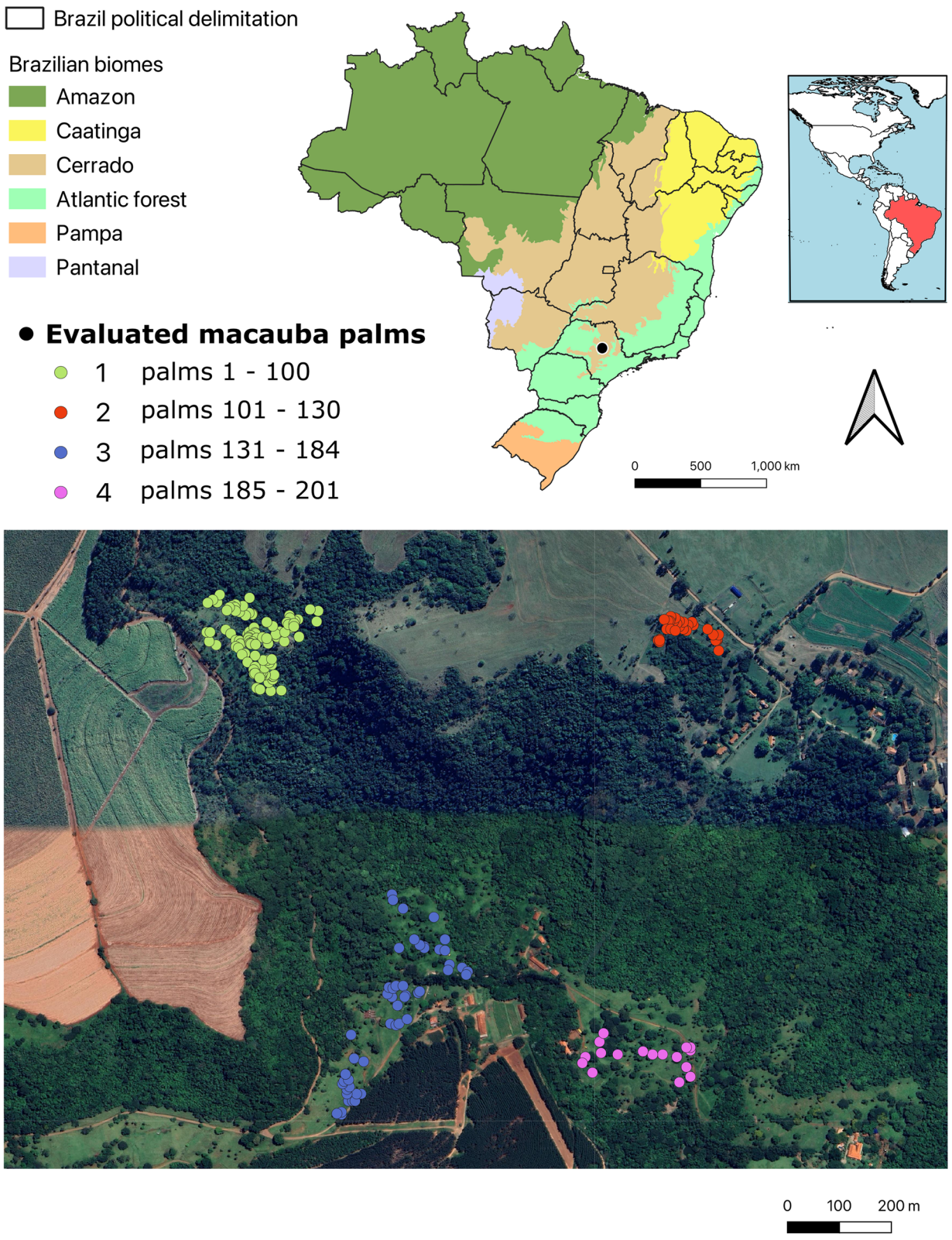

Fig. 1 Rural areas where macauba data were collected are shown. Green dots represent palms numbered 1 to 100, red dots represent palms 101 to 130, blue dots represent palms 131 to 184, and purple dots represent palms 185 to 201

Fig. 2 Macauba palm and oil production traits. **A** Macauba in the field. **B** Fruit samples and materials used to obtain the biometric fruit traits (hammer for breaking the fruits, scalpel for cutting the pulp and caliper for measuring the fruits). **C** Fruit fractions in the repetitions used to evaluate the biometric fruit traits; from top to bottom: husk, pulp, endocarp, and kernel. **D** Pulp dry mass macerated in the analytical mill IKA–A11. **E** Pulp dry mass samples ready for oil content quantification at NIR

leaf needles length (NL in centimeters) and leaf needle width (NW in centimeters). To obtain the H of the palms, a hypsometer was used. The DAP was measured at 1.5 m above the ground with a ruler. The LN was measured in units by counting the leaves of each plant, while the NN was measured by counting all the needles on the right side of the leaf. The LL was measured using a measuring tape, and the NL and NW were measured using a ruler.

The oil production traits encompass the biometric fruit traits and pulp oil content. Fruit characterization was based on six fruits from each palm [47] across three repetitions (Fig. 2B). We used two fruits per repetition to obtain a sufficient amount of pulp mass to evaluate the oil content. Therefore, we used the mean of the three repetitions in the biometric data and oil content, over two evaluation years (production seasons of 2019/2020 and 2020/2021) to perform the statistical analysis. The total fruit mass (FM in grams) was determined using a precision scale. After obtaining the total weight of the fruits, they were manually separated into four fractions: (1) husk, (2) pulp, (3) endocarp, and (4) kernel for fresh and dry mass (Fig. 2C). The fresh masses of the fruit fractions were dried in a ventilated oven at 36°C for 36 h. To study the biometrics fruit traits, we identified these traits as: husk fresh mass (HFM), pulp fresh mass (PFM), endocarp fresh mass (EFM), kernel fresh mass (KFM), husk dry mass (HDM), pulp dry mass (PDM), endocarp dry mass (EDM) and kernel dry mass (KDM). The weight of the fresh and dry masses of the fractions was also determined using a precision scale, in grams. The weights of the four dry fractions were added to obtain the value of the total mass of the dry fruit (FDM). Pulp oil content (OC, % or g/100 g) was quantified from the PDM (Fig. 2D and E) using Near-infrared Spectroscopy (NIRS). The PDM of the three replicates was crushed in the analytical mil IKA–A11. Diffuse reflectance of the samples was measured in rectangular cells (48 × 58 mm) using a FOSS NIRSystems 6500 spectrophotometer. The spectra measurements of the PDM were performed in triplicates and recorded using the ISIscanTM software, VERSION 3.10 (Infrasoft International, 2007). A multivariate model was built using the software Pirouette 4.5 (Infometrix, 1990–2011) to predict the oil content using the NIRs spectra.

Statistical analysis of the phenotypic data was performed using a mixed linear model in the lme4 package [48] in the R software, version 4.0.5 [49]. The boxplots displaying the mean, maximum, and minimum values of each trait by year are provided in the Supplementary Material. The statistical model used to analyse the phenotypic data is as follows:

$$Y_{ij} = \mu + g_i + y_j + \varepsilon_{ij} \quad (1)$$

where Y_{ij} is the phenotype of the *ith* palm in the *jth* year, μ is the general mean, g_i is the random effect of the palm i, y_j is the fixed effect of years j, and ε_{ij} is the residual.

The significance of the random effect of the palms was estimated by the likelihood ratio test (LRT) at 5% probability. The genetic ($\sigma^2{}_g$) and error $\sigma^2 e$ variance components were used to calculate the phenotypic variance $\sigma^2 p$:

$$\sigma^2{}_p = \sigma^2{}_g + \sigma^2{}_e \quad (2)$$

The coefficient of variation was calculated following the equation:

$$\mathrm{CV} = \frac{\sqrt{\sigma^2}\mathrm{e}}{mean} \times 100 \quad (3)$$

The variance components were used to estimate the broad sense heritability (H^2):

$$H^2 = \frac{\sigma^2{}_g}{\sigma^2{}_g + \frac{\sigma^2 e}{r}} \quad (4)$$

where r is the number of repetitions, which in this context corresponds to the number of evaluated years. All estimates below were computed using R software.

Genotypic correlation was calculated with multivariate linear mixed model using restricted maximum likelihood methods (REML) in the sommer R package, version 4.1.5. [50].

Genotype data

Genotyping was performed with markers based on single-nucleotide polymorphism (SNPs), following the protocol of GBS using two restriction enzymes (ddGBS) [51]. The genomic DNA was isolated from leaf material using the Doyle & Doyle protocol [52]. We evaluated the extracted DNA quality and quantity by agarose gel electrophoresis (1% w/v) stained with SYBR Safe DNA Gel Stain (Invitrogen) and by visual comparison with lambda DNA (Invitrogen). The quantification and normalization of genomic DNA were performed through fluorescence using the Qubit dsDNA BR Assay (Qubit-Life Technologies). Based on the obtained reading, we standardized the DNA to a concentration of 30 ng. μl^{-1}.

To obtain the SNPs, we prepared three 96-plex genomic libraries using the ddGBS technique according to the protocol described by [53] and the modifications used by Díaz et al. [20]. We digested the genomic DNA with the combination of enzymes *NsiI* and *Mse1* (New England Biolabs). The ddGBS libraries were quantified through RT-PCR on the CFX 384 Touch Real-Time PCR (Bio-Rad) equipment using a KAPA Library Quantification kit (KAPA Biosystems, cat. KK4824), and the fragments' profiles were inspected using the Agilent DNA 1000 Kit on a 2100 Bioanalyzer (Agilent Technologies). The 201 prepared sample libraries were sequenced on a single run in an Illumina HiSeq3000 with single-end and 101 bp configurations. The overall quality of the sequencing of GBS libraries was evaluated with the FastQC program [54]. Quality control and *demultiplex* were performed with the *process_radtags* module of the Stacks 1.42 program [55], where low-quality reads were removed. All the sequences were truncated to 90 bp due to a drop in overall mean sequence quality in FastQC towards the end of the raw sequences. After demultiplexing, five individuals were removed due to low quality and fewer than 20,000 reads retained in the sequencing, resulting in a total of 196 individuals.

The SNP calling was performed using three different strategies because no reference genome is available for macauba. First, (i) using the de novo pipeline (Stacks v.1.42) [55] based on the alignment of the reads obtained during the genotyping, (ii) using the genome of *Elaeis guineensis* var *tenera* (oil palm) as a reference [56], and (iii) using the transcriptome of *Acrocomia aculeata* as a reference [57]. SNP calling from the de novo pipeline was carried out using the software Stacks *v*. 1.42 [55]. After quality control, samples were demultiplexed using process-rad-tags module in Stacks. Then the *ustacks* module was used to identify groups of putatively homologous reads (putative loci) for each sample separately following the parameters: minimum sequencing depth (m = 3) and maximum mismatches (M = 2). Loci with lower values of probability (lnl_lim -10) were eliminated by the *rxstacks* correction module. The SNPs were filtered using the *populations* module, retaining all SNP in the same sequenced tag that had passed the filtering criteria.

For SNP calling using reference sequences, the *bwa-mem* algorithm of the *bwa* 0.7.17 program [58] was employed to align the sequences of each sample to the genome of *Elaeis guineensis* var *tenera* (oil palm) EG5 (NCBI GCA_000442705.1) [56] and to the transcriptome sequencing of *Acrocomia aculeata* [57]. Alignment files were processed with SAMtools [59] and Picard programs (http://broadinstitute.github.io/picard). SNP

identification was performed using the program *freebayes* 1.3.4 [60] with the configuration –standard_filters. VCFtools 0.1.17 [61] and bcftools 0.1.12 [59] programs were used to filter SNP markers, retaining all SNP per sequence in the same sequenced tag that had passed the filtering criteria.

To filter SNPs in the three SNP calling strategies, we used the following criteria: maximum number of alleles = 2, minor allele frequency ≥ 0.01, sequencing depth ≥ 3X, mapping quality ≥ 20, maximum percentage of 30% of missing data per locus and of 45% of missing data per individuals. Since we are working with a non-domesticated species like macauba, we opted for a more conservative approach in selecting the threshold for missing data per individual to avoid inaccuracies and bias in our analyses. After filtering, we identified a total of 27,410 SNPs in 153 individuals for the de novo genotypic dataset, 10,444 SNPs in 158 individuals using the oil palm genotypic dataset, and 4,329 SNPs in 167 individuals using the transcriptome genotypic dataset. Missing data were imputed using the Beagle 5.3 software [62, 63].

Chromosome information was added to the three genotypic datasets to perform GWAS analyses. For de novo and transcriptome genotypic datasets, which do not have a reference genome, we considered that all the SNP markers belonged to the same chromosome. For the oil palm genotypic dataset, chromosome information was obtained from the NCBI GCA_000442705.1. The oil palm reference genome contains information from 16 chromosomes and sequences that have not yet been allocated in the oil palm genome, referred as "SNW".

Genetic diversity and population structure

Genetic diversity was analyzed individually for each genotypic dataset, considering all the individuals within each dataset. For this purpose, two R packages were utilized: hierfstat [64] and adegenet [65]. In this analysis four macauba groups were considered due to the geographic distribution in nature (Fig. 1): group 1 (palms numbered 1 to 100), group 2 (palms 101 to 130), group 3 (palms 131 to 184) and group 4 (palms 185 to 201). Genetic diversity was investigated according to the value of the total number of alleles, the observed and expected heterozygosity, and the inbreeding coefficient. Meanwhile, population structure was inferred by discriminant analysis of principal components (DAPC). The DAPC was carried out through the poppr package to describe the evidence for individual cluster assignment of the palms in each genotypic dataset. The number of population cluster was obtained using the find.clusters function, with K-means set to 2. The individual group memberships from the DAPC were exploited using the posterior membership probabilities, based on the retained discriminant functions.

Single-trait and multi-trait GWAS

To perform the single-trait and multi-trait GWAS, we used the adjusted mean of the phenotypic data obtained from Eq. (1) in the statistical analyses, along with the three genotypic datasets. To obtaind the adjusted means, the genotype effect was considered fixed in the model. Thus, to conduct single-trait GWAS, four different statistical models were used and compared. These models, fitted using the GAPIT package (version 3) [66], were: (i) general linear model (GLM) [67], (ii) multiple loci MLM (MLMM) [68], (iii) fixed and random model circulating probability unification (FarmCPU) [40], and (iv) bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK) [41]. To account for population structure, we utilized the "Model.selection = TRUE" and "PCA.total = 5", which selects the best number of principal components from 0 to 5 in the GLM model. The relationship matrix was also calculated by GAPIT using its default parameters. Each model used in the single-trait GWAS has its own characteristics, so the use of population structure or kinship matrix depends on the model employed at the time of analysis.

The multi-trait analysis was fitted using a multivariate stepwise method (MSTEP) implemented in the software TASSEL, conducted in all pairwise trait combinations [46]. The population structure used for the single-trait analysis was also used in the multi-trait analysis. Prior to running MSTEP, each trait was normal quantile transformed using the orderNorm function from the R package bestNormalize [69]. Next, multivariate outliers were removed based on the aq.plot function from the R package mvoutlier [70]. In all cases, MSTEP was fitted with and without the first five principal components as covariates obtained from the respective marker data set. SNPs that were significant in both instances were selected as high-confidence multivariate associations. In summary, 15 analyses were performed in total: 12 in the single-trait GWAS and 3 in the multi-trait GWAS (5 models × 3 datasets).

To identify the candidate genes, tag sequences containing the significant SNPs from the de novo genotypic dataset were manually recovered from the catalog generated by the *cstacks* module in the Stacks program. The sequences of the significant SNPs from *Acrocomia aculeata* transcripts were obtained by retrieving fasta sequences from the transcriptome assembly with the perl fasta_FetchSeqs.pl script (https://github.com/4ureliek/Fasta). The occurrence of significant SNPs in predicted gene regions for *Elaeis guineenses* was verified with the *intersect* function from BEDTools v.2.30 program

[71], while the respective predicted protein sequences were recovered with the perl fasta_FetchSeqs.pl script. Blast2GO software [72] was used to search for similarities between the sequences of significant SNPs and candidate genes. The *blastx* function was used for the de novo genotypic dataset, and the *blastp* for the transcriptome and oil palm genotypic dataset. Venn diagrams were obtained from the site InteractiVenn [73].

Results

Phenotypic analysis

Mean phenotypic values and their range, estimates of σ^2_g, σ^2_p, and σ^2_e, CV, H^2, and LRT are shown in Table 1. The LRT detected significant differences among genotypes for all traits ($p < 0.001$) (Table 1). Mean values for the vegetative and oil production traits studied showed a wide range, indicating that this population has promising genotypes for selection and the initiation of breeding. In the Supplementary Material, there are density histograms and boxplots showing the distribution of the means for the traits evaluated in this study. The boxplots reveal an overlap of values and means between the evaluated years for most of the traits studied. Therefore, due to the lack of replication in the experiment, we decided to use the mean of the two years of evaluation in the phenotypic analyses rather than considering each year separately. This approach provides a more reliable estimate of the mean.

The CV ranged from 22.38% in LN to 5.75% in LL. Additionally, the traits H, LL, OC, FM, PFM, EFM, HDM, EDM, and KDM exhibited heritability values greater than 80%. FM had the highest genetic, phenotypic, and residual variance values (176.20, 248.20, and 72.00, respectively), while LL had the lowest values (0.05, 0.07, and 0.02, respectively). PFM, FDM, and OC also displayed high genetic variance. In the context of plant breeding, high genetic variance is advantageous as it provides more opportunities to select individuals with desirable traits, potentially leading to greater genetic gains in future generations. When selecting superior genotypes, the breeder should consider the commercial focus for the species, given that as macauba can be used for various purposes. For instance, the percentage of OC ranged from 17.47% to 60.80%, suggesting that superior genotypes for this valuable trait could be selected within this population for the biofuel and food indutries. Relevant oil production traits for the biofuel industry include FM, PDM, KDM, and OC. Therefore, by selecting the 20 trees with the highest averages from the studied population, they would represent the top 10% of individuals in the selection. Consequently, the selected individuals would have FM values above 95 g, PFM values exceeding 51 g, KFM values above 4.8 g, and OC values over 45% (data not shown). Conversely, if the breeder's goal is to use the endocarp for charcoal production, the selection of

Table 1 Means (range), estimates of genetic, phenotypic, and residual variances (σ^2_g, σ^2_p, and σ^2e, respectively), coefficient of variation (CV%), heritability (h^2) and likelihood ratio test (LRT) for genotypes in vegetative and oil production traits

Categories	Traits	Mean (range)	σ^2_g	σ^2_p	σ^2_e	CV%	H^2	LRT
Vegetative traits	H	8.26 (2.80—15.70)	3.39	3.71	0.31	6.84	0.95	346.55[a]
	DHB	28.77 (17.70—49.60)	9.28	17.40	8.12	9.90	0.69	65.75[a]
	LN	22.17 (6.00—42.00)	12.46	37.08	24.62	22.38	0.50	23.37[a]
	LL	2.68 (1.90—3.94)	0.05	0.07	0.02	5.75	0.81	124.87[a]
	NN	35.03 (22.00 – 52.00)	7.11	19.74	12.62	10.14	0.52	27.16[a]
	NL	6.79 (3.87 – 9.40)	33.59	66.86	33.27	8.49	0.66	56.56[a]
	NW	2.14 (1.23—3.47)	0.07	0.15	0.08	13.58	0.63	49.43[a]
Oil production traits	FM	76.51 (39.90—147.53)	176.20	248.20	72.00	11.09	0.83	98.61[a]
	HFM	16.98 (8.13—38.37)	11.34	17.24	5.89	14.30	0.79	74.92[a]
	PFM	38.71 (16.00—89.40)	74.24	108.22	33.98	15.05	0.81	87.42[a]
	EFM	14.76 (9.46—27.49)	5.69	7.79	2.10	9.82	0.84	114.21[a]
	KFM	3.74 (2.03—7.03)	0.51	0.79	0.28	14.17	0.78	74.60[a]
	FDM	45.26 (25.01—92.67)	58.03	90.02	31.99	12.49	0.78	76.04[a]
	HDM	9.96 (4.86—18.65)	3.76	5.50	1.74	13.25	0.81	93.91[a]
	PDM	19.74 (9.00—51.19)	21.52	32.81	11.29	17.02	0.79	74.00[a]
	EDM	12.47 (4.40—23.68)	4.22	5.70	1.48	9.75	0.85	118.52[a]
	KDM	3.01 (0.99—5.16)	0.34	0.51	0.16	13.66	0.80	79.41[a]
	OC	36.58 (17.47—60.80)	46.18	57.51	11.33	9.20	0.89	106.24[a]

[a] Significant according to the χ2 test (α = 0.01)

individuals should prioritize fruits with a higher EDM weight.

Pairwise genetic correlation between the vegetative and oil production traits studied are shown in Fig. 3. The oil production traits FM and FDM (0.94), HFM and HDM (0.90), PFM and PDM (0.91), EFM and EDM (1.00), KFM and KDM (0.99) had positive genetic correlation. Other traits that presented positive correlations were FM and HFM (0.83), FM and PFM (0.95), FDM and PFM (0.88), FM and PDM (0.86), FDM and HDM (0.81), FDM and PDM (0.92), indicating that these traits may be controlled by the same genes. This finding is valuable for breeders, as it implies that focusing selection on just one of these traits could streamline the biometric evaluation process of the fruit. For OC and NN, null and negative values were observed for all the pairwise traits tested. For example, NN and LL showed -0.39, while for NL and NN, the correlation was -0.47. OC had a correlation varying from 0.00 to -0.42 between the pairwise evaluated traits, indicating that the genes controlling these traits are distinct (Fig. 3).

Genetic diversity and population structure

Genetic diversity parameters from the macauba population studied showed different values for each SNP calling strategy (Table 2). In our study, the number of individuals and SNP markers used in the genetic diversity analyses varied in each genotypic dataset due to the SNP calling and quality control step. For the de novo genotypic dataset, the H_o and *Fis* were the same, at 0.15. For the oil palm genotypic dataset, the H_o was 0.41, while the *Fis* was -0.35. The transcriptome genotypic dataset showed values of 0.53 and -0.53 for H_o and *Fis*, respectively.

DAPC analysis of the three genotypic datasets revealed that the macauba individuals studied in this work belong to the same population (Fig. 4). The group 1 presented their individuals membership in two clusters, while groups 2, 3 and 4 showed their individuals membership in one cluster. The three genotypic datasets showed similar results to the group membership probabilities based on DAPC analysis. In group 3, only individual 175 exhibited 89%, 2.9%

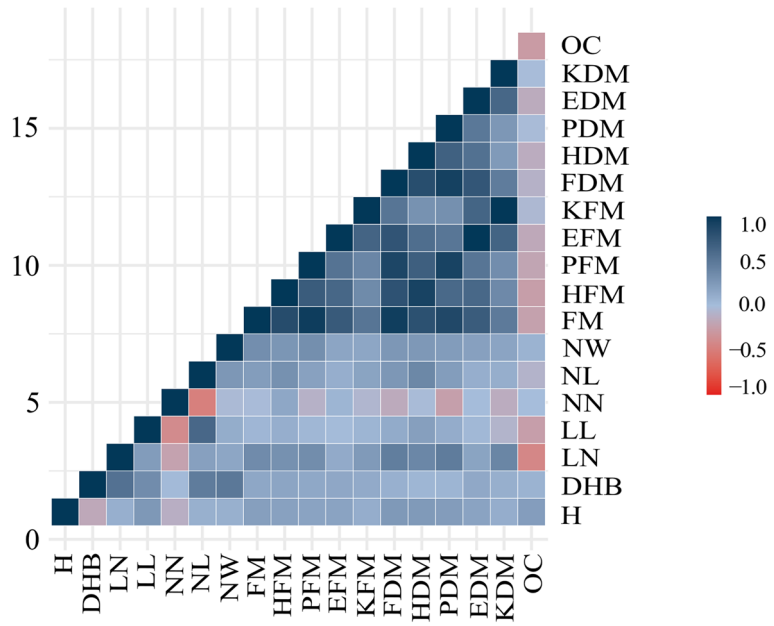

Fig. 3 Genetic correlation among vegetative and oil production macauba traits

Table 2 Population genetic parameters in macauba individuals: observed heterozygosity (H_o), expected heterozygosity (H_e), overall gene diversity (H_t), Wright's inbreeding coefficient (*Fis*)

Genotypic dataset	Number of individuals	Number of SNP	H_o	H_e	H_t	*Fis*
De novo	153	27,410	0.15	0.17	0.18	0.15
Oil palm	158	10,444	0.41	0.30	0.31	-0.35
Transcriptome	167	4329	0.53	0.34	0.34	-0.53

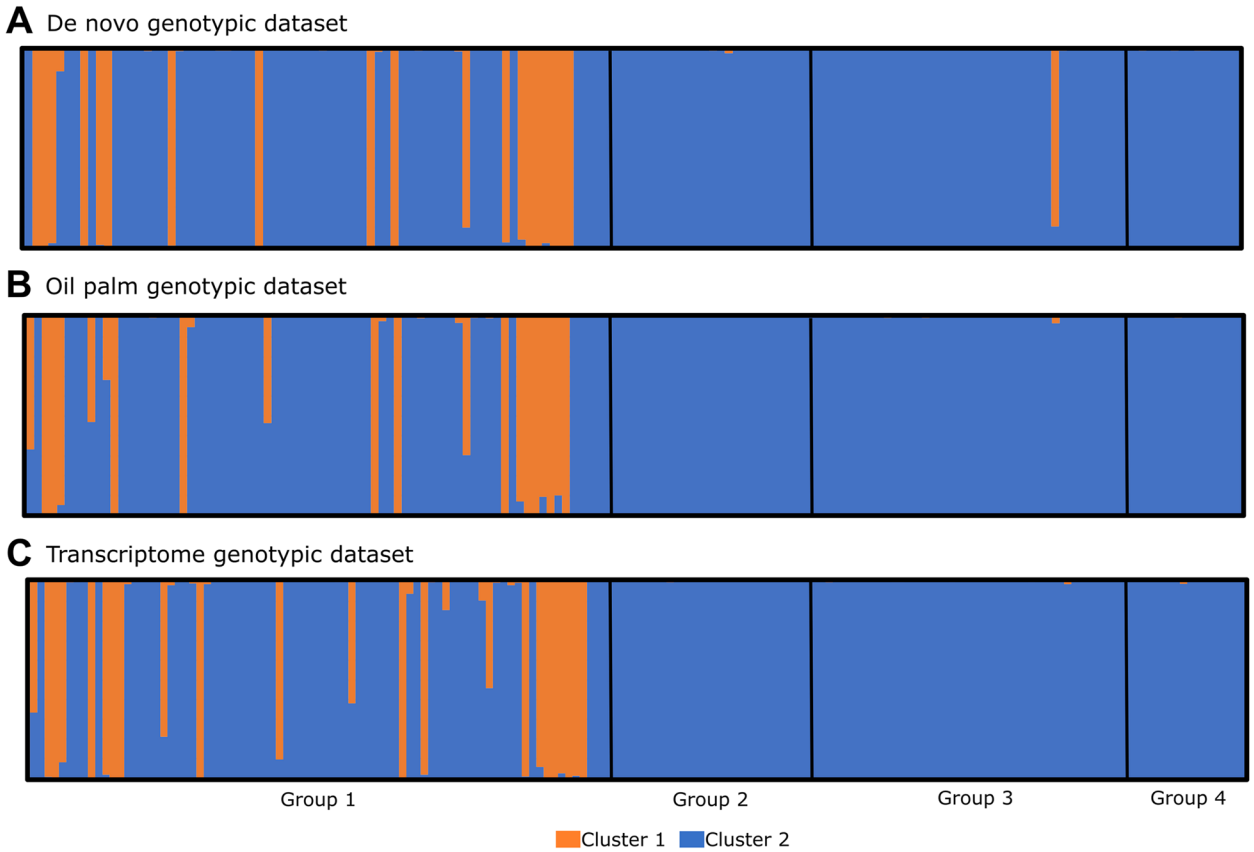

Fig. 4 Individual group membership probabilities based on DAPC analysis of genotypic datasets de novo (**A**), oil palm (**B**) and transcriptome (**C**) (k = 2). The horizontal line represents individuals from the four groups, showing their membership in the two clusters identified by orange and blue colors, with the probabilities of group membership indicated on the vertical axis

and 1.3% membership to the orange cluster in the de novo, oil palm and transcriptome genotypic datasets, respectively.

Single-trait and multi-trait GWAS

In the single-trait GWAS, a total of 112 significant SNPs and 13 candidate genes were identified (Supplementary Table 1). Of the 112 significant SNPs, 93 are associated with the oil production traits, while 19 are associated to the vegetative traits. (Supplementary Table 2). The three genotypic datasets used in this study showed significant SNPs for the GLM, MLMM, FarmCPU, and BLINK models. In the multi-trait GWAS, a total of 19 significant SNPs and four candidate genes were identified (Supplementary Table 1). The MSTEP model detected SNPs associated with traits in all the genotypic datasets. The significant SNPs detected from the single-trait and multi-trait GWAS can be found in the Supplementary Tables 1 and 2. The de novo genotypic dataset revealed a total of 92 significant SNPs associated with the DBH, NL, NW, LL, OC, FM, HFM, PFM, EFM, KFM, FDM, HDM, PDM, and EDM traits in all single-trait models implemented. The SNP markers 282174_18 and 271071_41 associated with the FM trait showed the highest and lowest effect value of 8.24 and -21.02 respectively. In these genotypic datasets, many of the significant SNPs were associated more than once in the same statistical model for different traits (Supplementary Table 3). The oil palm genotypic dataset showed 19 significant SNPs in the NW, FM, PFM, PDM and OC traits in the single-trait models GLM, MLMM, FarmCPU, and BLINK. The SNP marker SNC_025995.1 in the FM trait showed the highest effect value of 15.44. The SNP markers SNC_026005.1 and SNC_025995.1, associated with the PFM and OC, showed the lowest effect values of -12.67. The transcriptome genotypic dataset showed one significant SNP (STRINITY31179_1001) associated with the FDM trait in the MLMM model, with an effect of 55.96 (Supplementary Table 2). In our results we observed that different GWAS models detected the same significant SNPs

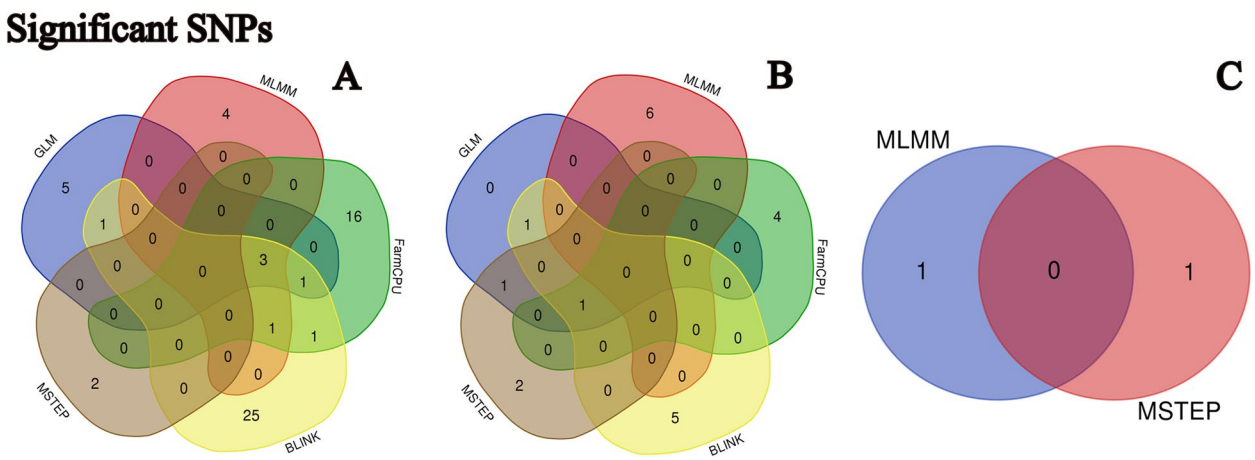

Fig. 5 Significant SNPs detected by different single-trait (GLM, MLMM, FarmCPU, BLINK) and multi-trait (MSTEP) GWAS models in each genotypic dataset are represented in a Venn Diagram. The diagram illustrates the unique and shared significant SNP in the de novo (**A**), oil palm (**B**), and transcriptome (**C**) genotypic datasets. The figure represents the total number of SNP markers in each genotypic dataset, excluding duplicated SNPs from the de novo genotypic dataset

inside the genotypic datasets. Fig. 5 represents the total number of shared significant SNPs across the GWAS models implemented in each genotypic dataset. In the same way, within the context of multi-trait GWAS, the association of the same significant SNP was observed across different combinations of traits in each genotypic dataset (Supplementary Table 4). For example, in the de novo genotypic dataset, the marker 333769_9 showed significant association with the traits HDM-PDM, FDM-OC, and FDM-HDM. In the oil palm genotypic dataset, the marker SNC_026000.1 was associated with the traits LN-HFM, NN-HFM, NN-FM, and NN-HDM. Meanwhile, in the transcriptome genotypic dataset, the marker STRINITY9279_665 showed association with the traits FDM-OC and FDM-HDM. In the Supplementary Material, there are the Manhattan plots and QQ-plots of the single-trait GWAS. The SNPs significantly associated with the vegetative and oil production traits studied are those gray triangles that appear above the cutoff line in the Manhattan plot. The red squares within each Manhattan plot represent the significant SNPs detected by the multi-trait GWAS model.

Identification of candidate genes

Candidate genes were identified based on similarities between the sequences of significant SNPs detected in this GWAS study and genome sequences in other species. In the de novo genotypic dataset, out of the 92 significant SNPs identified in the single-trait GWAS, two occurred in genic regions, with effect values of -0.90 and -10.43 observed for the EFM and HDM traits (Table 3). The SNP 620475_46 was associated with 20 candidate genes in different species, all of them showing 100% similarity (data not shown). Table 3 presents three of these 20 candidate genes, which are similar to those in *Elaeis guineenses* and *Phoenix dactylifera*, and are involved in carbohydrate metabolic process. In the oil palm genotypic dataset from the single-trait GWAS, 10 significant SNPs associated with NW, OC, FM and PFM traits occurred in genic regions, with effect values ranging from -12.67 to 7.80 (Table 4). In the transcriptome genotypic dataset from the single-trait GWAS one significant SNP associated with FDM was annotated to a gene in *Elaeis guineensis* with an effect value of 55.96 (Table 3). In the multi-trait GWAS, significant SNPs showed similarities with candidate genes in 10 trait combinations (Table 6). In the oil palm genotypic dataset, the SNP markers SNC_025993.1_60355952, SNC_025995.1_17709887, and SNC_026000.1_5081018 had similarities with the candidate genes LOC105060459, LOC105041056 and LOC105049570, respectively. The SNP marker SNC_025993.1_60355952 associated with the trait combinations H-NW and DBH-NW, showed similarity with a candidate gene known for its membrane function (Table 5). This SNP marker was also associated with NW in three single-trait models (Table 4, Fig. 6). Furthermore, the SNP marker SNC_025995.1_17709887 which was associated with the trait combinations LL-OC and LN-OC, was also detected in the GLM single-trait model for OC, indicanting a metal ion binding function (Table 4, Fig. 6). For the transcriptome genotypic dataset, the marker STRINITY9279_4262 associated with the traits FDM-OC and FDM-HDM showed similarity with an uncharacterized protein (Table 5). In all the genotypic datasets used in this GWAS study, candidate genes showed similarity with *Elaeis guineensis* (100%), *Phoenix*

Table 3 Candidate genes from the de novo and transcriptome genotypic dataset significantly associated with EFM, HDM and FDM traits as identified by the BLINK, GLM and MLMM single-trait GWAS models

Dataset	Trait	Single-trait GWAS	SNP marker	Candidate gene	Chr	Position (bp)	*P* value	MAF	Effect	Gene Annotation	Scientific Taxonomy	Function
De novo	EFM	BLINK	568099_49	EHA8587225.1	UN	39,766,911	0.0000001	0.38	-0.9	Hypothetical protein gb|EHA8587225.1|	*Cocos nucifera*	
	HDM	GLM	620475_46	LOC105052613	UN	43,433,228	0.0000021	0.05	-10.43	Protein reduced wall acetylation 4 isoform X1 and X2	*Elaeis guineensis*	Carbohydrate metabolic process; Transferase activity; Golgi apparatus; Membrane; Membrane
				LOC105039974	UN	43,433,229	0.0000021	0.05		Protein reduced wall acetylation 4	*Elaeis guineensis*	
				LOC103719307	UN	43,433,230	0.0000021	0.05		Protein reduced wall acetylation 4-like isoform X1, X2, X3, X4 e X5	*Phoenix dactylifera*	
Transcriptome	FDM	MLMM	STRINITY31179	LOC105037057	1	1001	0.0000045	0.49	55.96	Uncharacterized protein LOC105037057	*Elaeis guineensis*	DNA repair; DNA binding; Catalytic activity; Nucleus

Table 4 Candidate genes from the oil palm genotypic dataset significantly associated with NW, OC, FM, and PFM traits as identified by various single-trait GWAS models

Trait	Single-trait GWAS	SNP marker	Candidate gene	Chr	Position (bp)	*P* value	MAF	Effect	Gene Annotation	Scientific Taxonomy	Function
NW	GLM	SNC_025993.1_60355952	LOC105060459	1	60,355,952	0.0000029	0.44	0.18	Uncharacterized protein	*Elaeis guineensis*	Membrane
	FarmCPU					0.0000000		0.14			
	BLINK					0.0000000		0.17			
	BLINK	SNC_025997.1_18357165	LOC105045291	5	18,357,165	0.0000034	0.07	0.22	Ribosome biogenesis protein NSA2 homolog	*Elaeis guineensis*	rRNA processing
OC	GLM	SNC_025995.1_17709887	LOC105041056	3	17,709,887	0.0000023	0.02	-12.67	Zinc finger protein VAR3, chloroplastic	*Elaeis guineensis*	Metal ion binding
	FarmCPU	SNC_025999.1_11453414	LOC105048521	7	11,453,414	0.0000003	0.03	7.80	Uncharacterized protein	*Elaeis guineensis*	Protein transport
	FarmCPU	SNC_026004.1_15076493	LOC105054947	12	15,076,493	0.0000039	0.31	-2.61	Mitochondrial import inner membrane translocase subunit Tim13 isoform X2	*Elaeis guineensis*	Protein transport; Mitochondrial inner membrane;
FM	BLINK	SNW_011550986.1_491652	LOC105032941	-	-	0.0000018	0.18	-6.47	Glutathione reductase, cytosolic isoform X2	*Elaeis guineensis*	Glutathione metabolic process; Cell redox homeostasis; Cellular oxidant detoxification; Flavin adenine dinucleotide
	BLINK	SNW_011552849.1_78068	-	-	-	0.0000001	0.14	0.37	GDSL esterase/lipase LIP-4	*Elaeis guineensis*	Hydrolase activity
PFM	MLMM	SNC_025993.1_67158358	LOC105034069	1	67,158,358	0.0000007	0.08	-6.83	Extracellular ribonuclease LE-like	*Elaeis guineensis*	RNA phosphodiester bond hydrolysis, endonucleolytic; Ribonuclease T2 activity
	MLMM	SNC_025997.1_8305530	LOC105044859	5	8,305,530	0.0000019	0.18	4.91	Putative zinc transporter At3g08650	*Elaeis guineensis*	Metal ion transport; Transmembrane transport; Plasma membrane
	MLMM	SNC_026005.1_24632742	LOC105056693	13	24,632,742	0.0000003	0.02	-12.67	20 kDa chaperonin, chloroplastic	*Elaeis guineensis*	Protein folding; Positive regulation of superoxide dismutase activity; ATP binding;

Table 5 Candidate genes from the oil palm and transcriptome genotypic dataset significantly associated with different combinations of traits identified by the MSTEP Multi-trait GWAS model

Genotypic dataset	Trait	SNP marker	Candidate gene	Chr	Position (bp)	ProbF	Gene Annotation	Scientific Taxonomy	Function
Oil Palm	H-NW	SNC_025993.1_ 60,355,952	LOC105060459	1	60,355,952	0.0000008	Uncharacterized protein	*Elaeis guineensis*	Membrane
	DBH-NW					0.0000084			
	LL-OC	SNC_025995.1_17709887	LOC105041056	3	17,709,887	0.0000003	Zinc finger protein VAR3, chloroplastic	*Elaeis guineensis*	Metal ion binding
	LN-OC					0.0000012			
	LN-HFM	SNC_026000.1_55081018	LOC105049570	8	5,081,018	0.0000053	UDP-N-acetylglucosamine peptide N-acetylglucosaminyltransferase 110 kDa subunit isoform X1	*Elaeis guineensis*	-
	NN-HFM					0.0000000			
	NN-FM					0.0000001			
	NN-HDM					0.0000019			
Transcriptome	FDM-OC	STRINITY9279_665	LOC105039939	1	665	0.0000011	Uncharacterized protein	*Elaeis guineensis*	-
Transcriptome	FDM—HDM					0.0000412			

dactylifera (100%), and *Cocos nucifera* (91.3%). In general, they act in protein transport, DNA repair, metal ion transport, maturation of ribosome subunit and transmembrane transportation.

Discussion

Acrocomia is a genus of neotropical palms distributed across almost all tropical and subtropical regions of the Americas. There are nine recognized species in this genus: *A. aculeata*, *A. totai*, *A. hassleri*, *A. glaucescens*., *A. emensis*; *A. intumescens*; *A. media, A. crispa,* and *A. corumbaensis* [1, 3, 20]. Among these, macauba (*Acrocomia aculeata*) has gained prominence in recent years due to its high oil content in both the pulp and kernels of its fruit, with productivity comparable to that of the oil palm. To our knowledge, this is the first GWAS study conducted on *Acrocomia aculeata*. Understanding the genetic architecture of macauba is a crucial step towards its domestication and commercial cultivation, as identifying markers associated with economically important traits can aid in selecting promising genotypes for breeding programs.

Importance of the analyzed traits to *Acrocomia aculeata* domestication

The data used in this study were collected from a natural population on a private rural property. Analyzing macauba phenotypic data help us understand the relationships between traits and guides future breeding strategies for the species. This study revealed significant genetic effects, indicanting substancial genetic variability within the evaluated macauba population (Table 1). We also observed a wide range of phenotypic variation. Since macauba is an incipiently domesticated palm, this variation results from two factors: genetic and environmental effects, and the fact that we sampled a native population. The lack of an experimental design means that the age of the macauba genotypes was unknown. For instance, traits like H and DHB could not be consistently obtained across plants of the same age. Therefore, it is important to consider that the difference in height of a genotype compared to others may be due to age as well as genetic merit. In this study, we were unable to distinguish this difference. This is one of the major challenges in studying vegetative traits in perennial populations in their natural environment. In macauba, flowering typically begins in the fifth or sixth year of the palm's life [74], and fruit development is slow, lasting up to 62 weeks after anthesis, or approximately 14.3 months [75, 76]. Additionally, the trees can develop six or more fruit bunches, each with a distinct maturation cycle. As a result, genotypes exhibit fruit ripening from October to April, with ripe fruits undergoing abscission and falling to the ground [76]. In our research, during field collection, we selected trees that had recently shed fruits and whose fruits showed clear signs of ripening, aiming to standardize mature fruits for our study.

Even without an experimental design, the CV ranged from 5.75% to 22.38%, indicating good accuracy in the phenotypic data obtained. Additionally, traits such as H, LL, OC, FM, PFM, EFM, HDM, EDM, and KDM exhibited heritability values ranging from 50 to 80%, demonstrating that breeding cycles in this population are likely to yield genetic improvements in these traits. Previous studies have also reported high mean heritability values for various morphological traits in macauba [23, 77, 78, 78].

Correlation between traits is highly relevant in plant breeding because it facilitates indirect selection of multiple traits [79, 80]. In this context, genotypic correlation is used in data analysis to guide breeding programs, as it is heritable [81]. Moreover, high values of genetic correlation between traits may indicate the presence of linked genes or pleiotropy [42]. In this study, we

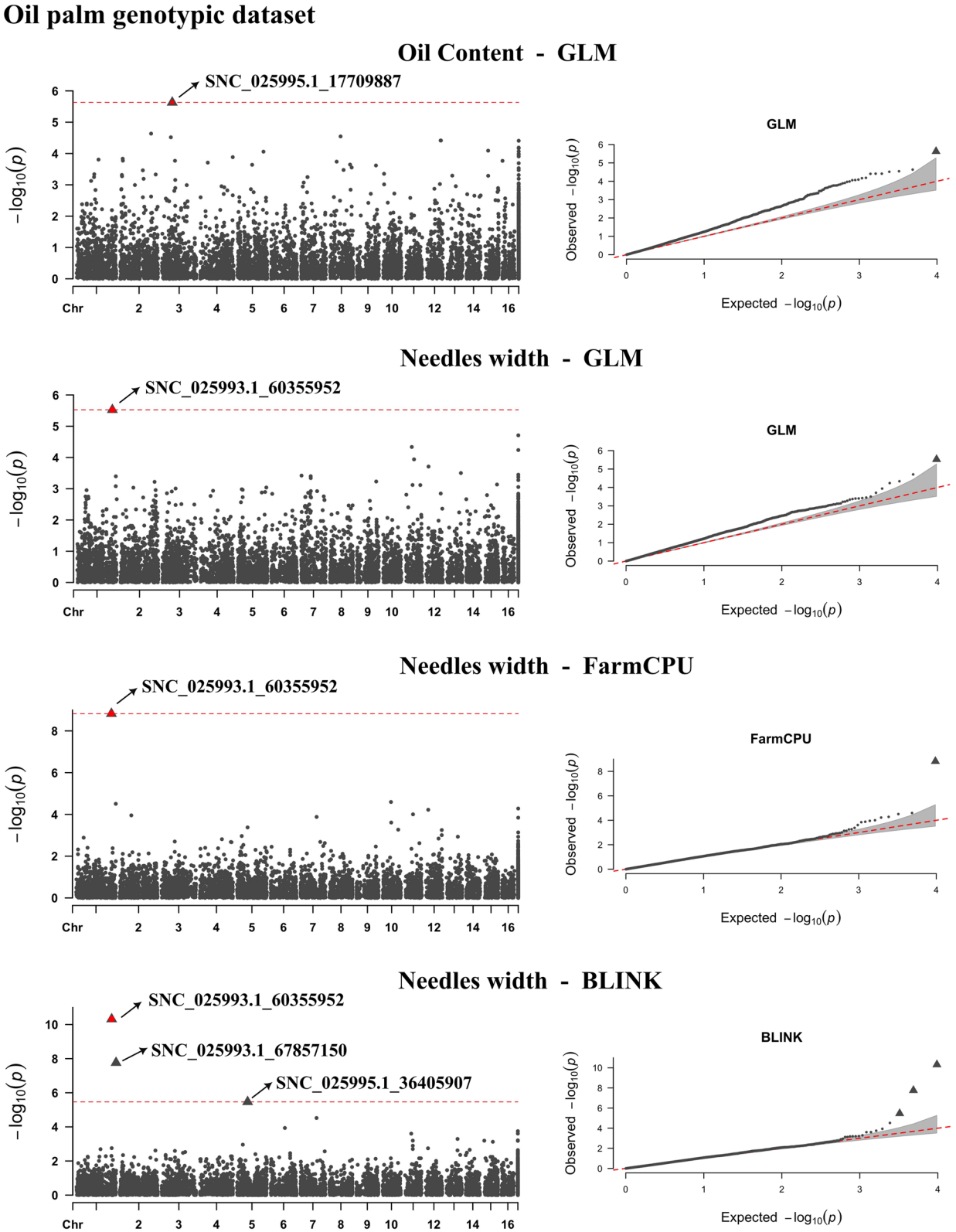

Fig. 6 Manhattan plots displaying SNPs detected both in single-trait and multi-trait GWAS from the oil palm genotypic dataset focusing on NW (SNC_025993.1_60355952 marker) and OC (SNC_025995.1_17709887 marker) traits. Significant SNPs identified by GLM, BLINK and FarmCPU are represented by gray triangle, while red squares represent the same SNPs detected in the multi-trait GWAS. The QQ plots corresponding to the models are to the right of each Manhattan plot

observed high genetic correlation values between the oil production traits such as FM-HFM, PFM-PDM, FDM-PFM, and HDM-PDM (Fig. 3). In the literature, other authors have reported strongly positive correlations among traits related to dry mass from oil production traits [8, 22, 76, 77], as well as negative and low correlation values [22]. Additionally, the FM-FDM, HFM-HDM, PFM-PDM, EFM-EDM, and KFM-KDM traits showed positive genetic correlations because they represent the same traits measured in different ways. The key difference between them is the presence of humidity. In the industry, all parts of the macauba fruit can be used for raw material production, and this use will depend on whether the mass is fresh or dry. For example, the fresh mass can be used for human and animal food, as well as by the pharmaceutical industry, while the dried mass can be used by bioenergy industry.

The statistical analysis presented in Table 1 and Fig. 3 provides guidance for selecting promising palms within the studied population based on commercially relevant traits, with the goal of initiating breeding efforts. As we are working with an orphan perennial species like macauba, we recommend that breeders initially focus on oil production traits, while vegetative traits should be considered secondary. This recommendation is supported by the high positive genetic correlation values observed for oil production traits and the fact that most significant SNPs detected in the GWAS analysis are associated with these traits. Consequently, future research linking genotype and phenotype could facilitate the selection of superior genotypes for oil production through association mapping.

The oil in macauba is found in both the pulp and the kernel, making both sources appealing to the energy industry. To select trees for biodiesel production, key traits of interest include FM, PDM, KDM, and OC. Higher values for these traits increase their attractiveness to the energy sector. Our observations reveal a positive correlation between fruit mass and pulp mass, indicating that heavier fruits contain more pulp. Consequently, selection efforts can concentrate on FM, facilitating a more efficient selection process, as fresh fruits can be weighed directly without the need for drying or desiccation to assess pulp mass. The OC demonstrated a weak negative correlation with FM, PDM and KDM. Therefore, the selection for OC should be conducted separately during the advanced stages of the breeding cycle. Initially, however, selection can be based on a combination of all these traits. In this context, we applied a rank-sum index to select the top individuals from the evaluated population, aiming to start a breeding cycle focused on increasing oil content for biodiesel production. Based on the simultaneous selection of the traits FM, PDM, KDM, and OC, the 15 top-performing individuals were 42, 17, 136, 54, 18, 49, 102, 77, 163, 19, 130, 43, 57, 7, and 164. Of these, individual 43 exhibited the highest OC value, reaching 61.19%, the highest among all 201 individuals evaluated. Considering only the OC trait in the selection, the individuals with oil content values above 50%, in descending order, were 43, 44, 45, 38, 152, 129, 55, and 52.

In the food industry, selecting superior individuals should focus on traits that exhibit high values for HFM, HDM, PFM, PDM, KFM, KDM, and OC. The fruit husk can be utilized for flour production, while the oils extracted from both the pulp and kernel can be incorporated into various industrial products. Moreover, the selection of HFM and HDM can also be informed by FM due to the positive correlation between these traits. In this case, simultaneous selection using the rank-sum index can focus on the traits HFM, PFM, KFM, and OC, since HFM-HDM, PFM-PDM, and KFM-KDM showed strong genetic correlations. Accordingly, the individuals selected for the food industry would be 42, 136, 163, 17, 54, 50, 49, 77, 164, 20, 191, 130, 102, 124, and 66. In these individuals, the oil content in the pulp ranged from 33.22% to 49.54%. Among vegetative traits, the number of leaves stands out as particularly noteworthy; in palms, each leaf axil develops into an inflorescence [82]. Thus, theoretically, a greater number of leaves corresponds to more inflorescences and, consequently, increased fruit production. In this selection, considering only the number of leaves on the tree, the five individuals with the highest leaf count would be 152, 135, 99, 126, and 77. Berton [77] proposed an ideotype model for improved macauba, which includes traits such as early flowering, low height, high production of fruits and oil content, fewer or no thorns, indehiscent fruits, and overlapping bunches. Therefore, studying macauba traits is essential to aid in selecting potential plants that exhibit the characteristics needed for the ideotype demanded by the industries.

Absence of a reference genome in *Acrocomia aculeata*

Because *Acrocomia aculeata* does not have a reference genome, different strategies were proposed by us to perform SNP calling. These strategies include using the de novo pipeline [55], the reference genome of *Elaeis guineensis* var *tenera*, and the transcriptome of *Acrocomia aculeata* [57]. The *Elaeis guineensis* reference genome was used in this work due to its phylogenetic proximity to *Acrocomia aculeata.* Both species belong to the subfamily Arecoideae and share morphological characteristics typical of palms in this subfamily [82, 83]. Bazzo et al., [57] used the *Elaeis guineensis* reference genome in their study, which identified and

validated 145 macauba EST-SSR markers from various tissues using transcriptome sequencing. The mRNA libraries were mapped against the *Elaeis guineensis* reference genome [56]. The cross-transferability of these EST-SSR marker to other palms showed a transferability rates of 80.7% in African oil palm. Moreover, recent phylogenomic studies using chloroplast genome sequences produced a phylogenomic tree in which all nodes had a posterior probability of 1.0 (PP = 1.0). In this phylogenomic tree, *Acrocomia aculeata* showed close proximity to *Elaeis guineensis* [84].

Population genetic parameters for macauba individuals from the oil palm and transcriptome genotypic datasets showed higher heterozygosity values compared to those from the de novo genotypic dataset (Table 2). Consequently, the inbreeding coefficient was higher in the population of the de novo genotypic dataset than in the other two datasets. Díaz et al. [20] investigated genetic diversity within the *Acrocomia* genus using genome-wide SNP and the de novo genotypic pipeline. They reported low heterozygosity values (0.031) for *Acrocomia aculeata*. Similarly, in our study, the population genetic parameters from the de novo genotypic dataset exhibited low heterozygosity values and higher inbreeding coefficient compared to the other genotypic datasets. These results are attributed to the use of the de novo pipeline in the Stacks software, where homologous reads were employed to identify SNPs [55]. In contrast, the oil palm and transcriptome datasets exhibited higher heterozygosity values and lower inbreeding coefficient values. This difference is likely due the *Elaeis guineensis* reference genome, which offers greater genetic variability in its alleles compared to the genome of the de novo pipeline. Although *Acrocomia* and *Elaeis* are closely related genera, they are expected to possess different alleles due to their distinct evolutionary history [82, 83]. The transcriptome dataset was generated by the alignment step with the transcriptome sequencing of various phenotypic traits of *Acrocomia aculeata* [57].

The DAPC analysis of the macauba population revealed that the individuals belong to a single population, which is divided into two distinct genetic clusters (Fig. 4). Given that groups 3 and 4 are geographically situated in lower altitude areas compared to groups 1 and 2, the admixture analysis suggests that groups 1 and 2 are the ancestors of groups 3 and 4. Individuals in group 1 exhibited both genetic clusters (orange and blue), whereas individuals in groups 2, 3, and 4 were predominantly from the blue cluster. The analysis of the genetic diversity in macauba is crucial for selecting the most promising materials for use, maximizing genetic gains, and more effectively contributing to the development of commercial cultivars [20]. The three genotypic datasets used in this study showed 0.15, 0.41 and 0.53 values of observed heterozygosity and 0.15, -0.35 and -0.53 values of inbreeding, indicating genetic variability in the population and accrediting it for breeding purposes.

Single-trait and multi-trait GWAS in macauba

Using both single-trait and multi-trait GWAS models, we identified SNP markers located in gene regions associated with vegetative and oil production traits in a macauba population. The results related to the genotype-to-phenotype association for these traits are significant for understanding the genetic architecture of this neotropical palm. Additionally, the genetic mapping and molecular characterization of genes contributing to the variation of complex traits could enhance genome-assisted breeding efforts crop improvement [85]. Unraveling this genetic information may also accelerate the implementation of breeding programs aimed at selecting superior genotypes for traits related to macauba oil production.

The total number of significant SNPs detected in the single-trait and multi-trait GWAS was greater than the number of candidate genes identified (Supplementary Table 1). These results are the first for *Acrocomia aculeata* and underscore the importance of obtaining a reference genome for the species. The single-trait GWAS identified significant genomic regions in all three datasets used in this study, associated with the following traits: DBH, LL, NL, NW, OC, FM, HFM, PFM, EFM, KFM, FDM, HDM, PDM and EDM. However, no genomic regions associated with the traits H and LN were identified in the similarity search using Blast2GO (Tables 4, 5, and 6).

We observed that different single-trait GWAS models detected the same loci (LOC105060459) for the same trait (Table 4), or the same SNP markers were identified for different trait combinations (Supplementary Table 2). Similar results were observed in the multi-trait model (Table 5). Moreover, in the multi-trait model, four candidate genes (SNC_025993.1, SNC_025995.1, SNC_026000.1, STRINITY9279) were notably associated with different combinations of traits. Among these traits, NW, OC, NN, HFM, HDM, and FDM appeared multiple times. Additionally, SNP markers SNC_025993.1 and SNC_025995.1, detected in the multi-trait model, were also identified in the single-trait models (Fig. 6). These results confirm the presence of these loci in the macauba genome, reducing the likelihood of these SNP markers being false positives. Fernandes et al. [42] suggest that using both single-trait and multi-trait GWAS is essencial to infer whether causal mutations underlying

peak GWAS associations are pleiotropic. However, they also emphasize that statistical analysis alone cannot distinguish between QTNs in linkage disequilibrium and a single pleiotropic QTN, underscoring the importance of validating significant SNPs detected by the GWAS.

Given the nature of our data, which originates from an experiment without a design, we also conducted single-trait GWAS for each year separately across the three genotypic datasets, and for both vegetative and oil production traits (Supplementary Table 5, 6, and 7). The number of significant SNPs identified was higher in the year-based GWAS compared to the GWAS using the adjusted mean, as presented in this manuscript. Since we are working with phenotypes from an unreplicated trial, the lack of correction in the mean likely results in a higher number of false positives, which could explain the larger number of significant SNPs in the year-based analysis. In this case, the significant SNPs that appear in both the year-based GWAS (for years 1 and 2) and the GWAS using the adjusted mean are more likely to represent true positive associations.

Candidate genes

In general, the candidate genes detected in this study were involved in processes such as RNA maturation, metal ion binding and transport, protein transportation, DNA repair, carbohydrate metabolic process, and other cell regulation biological processes.

From the single-trait GWAS the candidate gene LOC105045291, associated with the NW trait, encodes a polypeptide of 260 amino acids. This protein is annotated as a ribosome biogenesis protein NSA2 homolog (Table 4). NSA2 (Nop seven-associated 2) is a nucleolar protein linked to the ribonucleoprotein complex and plays a role in cell proliferation and cell cycle regulation. It was identified through high throughput screening of novel human genesand is evolutionarily conserved across different species [86]. For the oil content trait, three candidate genes were detected (Table 4). One of these genes, LOC105041056, encoded the Zinc finger protein VAR3 wich consist of 758 amino acids. This protein was also identified in the multi-trait GWAS for the trait combination LL-OC and LN-OC (Table 5). Zinc finger domain proteins are crucial for various cellular functions, including transcriptional regulation, RNA binding, apoptosis regulation, and protein–protein interactions [87]. Also, significant selection signatures (outlier SNPs) identified by Morales-Marroquín et al. [26] were associated with genes involved in fatty acid and triacylglycerol biosynthesis pathways in *Acrocomia aculeata*. This study detected signatures in genes related to zinc ion binding processes. Our findings suggest that candidate genes regulating oil production traits are linked to metal ion binding and related with traits such as OC, PFM, LL, and LN. A study in Arabidopsis using recessive variegated 3 (var3) mutants to investigate the VAR3 gene observed that var3 is part of a protein complex essential for normal chloroplast and palisade cell development [88].

For the FM trait, the SNP marker SNW_011552849.1_78086 was blasted against the GDSL esterase/lipase LIP-4 gene, which has hydrolase activity (Table 5). The GDSL esterase/lipase protein encompasses a variety of lipolytic enzymes that hydrolyze diverse lipidic substrates, including thioesters, aryl esters, and phospholipids [89]. Cao et al. [90] studied GDSL esterase/lipase in Rosaceae genomes and found that it plays a role in fruit development. In macauba, the fruiting is supra-annual, and the fruit growth curve follows a double sigmoidal trend with four stages, as observed by Montoya et al. [76]: the first stage features slow growth and negligible differentiation of the fruit's inner parts, the second stage includes the first growth spurt and differentiation of the inner parts; in the third stage, fruit growth slows, and all structures achieve differentiation; and finally, the second growth spurt and fruit maturation. Considering this, we can hypothesize that the GDSL esterase/lipase LIP-4 gene is involved in macauba fruit development. However, to validate this hypothesis, future research should be conducted to confirm this candidate gene in the species.

For the PFM trait, the 20 kDa chaperonin protein located in the chloroplast was detected by the MLMM single-trait model (Table 5). In general, molecular chaperones functionally support protein translocation across membranes, promote complex assembly and disassembly, and participate in many other regulatory processes within the cell [91–93]. In *Elaeis guineensis*, a study using proteomics and chemometrics approach provided important information about protein regulation during fruit ripening and oil synthesis [94]. In this study, the 20 kDa chaperonin was identified as a folding protein, and the authors observed that it had a down-regulated towards fruit ripening.

Candidate genes from the de novo genotypic dataset in the single-trait GWAS showed gene annotation for the Reduced Wall Acetylation 4 protein in the HDM trait (Table 3). The Reduced Wall Acetylation 4 protein is part of a family protein involved in the O-acetylation of cell wall polysaccharides in the Golgi apparatus. Specifically, the Reduced Wall Acetylation 4 is thought to be responsible for the translocation of acetyl-CoA across the Golgi membrane and appears to supply the acetyl-donor to both pectins and hemicelluloses [95, 96]. The polysaccharides pectins, hemicelluloses, and celluloses are the main cell wall components in the epidermis, contributing to its protection against xenobiotics, ultraviolet light, and pathogens and providing a waterproof barrier [97]. In this context, the Reduced Wall Acetylation 4 protein can be involved in the same cell functional route in the epidermis of the husk of macauba fruits.

From the multi-trait GWAS, the same candidate gene LOC105049570 was detected for the combination of the traits: LN-HFM, NN-HFM, NN-FM, and NN-HDM (Table 5). The protein associated with the SNC_026000.1_ 5,081,018 SNP marker is UDP-N-acetylglucosamine peptide N-acetylglucosaminyltransferase. In cucumber, a study identified and mapped the CsSF4 gene, which encodes for the protein mentioned above [98]. The authors observed that the CsSF4 gene was highly expressed in the leaves and male flowers of wild-type cucumbers and that it is required for fruit elongation and development. Since the UDP-N-acetylglucosamine peptide N-acetylglucosaminyltransferase protein was detected for different trait combinations, future studies aimed at validating the candidate gene LOC105049570 are necessary. Through them, it will be possible to confirm the presence of pleiotropy or genes linked to these traits.

Conclusions

In our study, we have identified novel genomic regions and candidate genes associated with vegetative and oil production traits in macauba. These candidate genes require further validation through targeted functional analyses, and multi-trait associations should be explored to determine whether pleiotropic or linked genes are present. This study represents the first application of GWAS in macauba, and the markers associated with key traits could serve as valuable tools for developing marker-assisted selection, contributing to macauba's domestication and pre-breeding efforts.

Abbreviations

H	Height
DHB	Diameter at breast height
LN	Number of leaves
LL	Leaves length
NN	Number of leaf needles
NL	Leaf needles length
NW	Leaf needles width
FM	Total fruit mass
HFM	Husk fresh mass
PFM	Pulp fresh mass
EFM	Endocarp fresh mass
KFM	Kernel fresh mass
FDM	Total mass of the dry fruit
HDM	Husk dry mass
PDM	Pulp dry mass
EDM	Endocarp dry mass
KDM	Kernel dry mass
OC	Pulp oil content
DAPC	Discriminat analysis of principal components
SNPs	Single-nucleotide polymorphism
GWAS	Genome-wide association studies
GLM	General linear model
MLMM	Multiple loci MLM
FarmCPU	Fixed and random model circulating probability unification
BLINK	Bayesian-information and linkage-disequilibrium iteratively nested keyway
MSTEP	Multivariate stepwise method
σ_g^2	Genetic variances
σ_P^2	Phenotypic variance
σ_e^2	Residual variance
CV	Coefficient of variation
H^2	Broad sense heritability
LRT	Likelihood ratio test
H_o	Observed heterozygosity
H_s	Expected heterozygosity
H_t	Overall gene diversity
Fis	Wright's inbreeding coefficient
MAF	Minor allele frequency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-05805-y.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.

Acknowledgements

We would like to express our gratitude to Luiz Henrique Berton, Gabriel Alves, Elivelton Alves, and Evandro Coelho from the Campinas Agronomic Institute, as well as Marcela Barbosa, Laecio Sampaio, and Marcelo Almeida from the Luiz de Queiroz College of Agriculture, for their invaluable assistance with phenotypic data collection and fruit biometry. Our thanks also go to Bárbara Regina Bazzo, Lucas Miguel de Carvalho, and Marcelo Falsarella Carazzolle, for providing the macauba transcripts used in the SNP calling procedures. Finally, we appreciate Saulo Fabrício das Silva Chaves for his suggestions during the manuscript review process.

Authors' contributions

EGOC participated in these stages of manuscript preparation: collected and processed phenotypic and genotypic data, developed the methodology, curated the data, conducted data mining analyses, and performed statistical analyses. JMM collected phenotypic data and implemented the methodology for obtaining genotypic data. AAP conducted bioinformatics analyses and data mining. SBF developed the methodology and conducted statistical analyses. CAC supported the investigation, collected phenotypic data, and provided financial resources. JAAF supported the investigation and collected phenotypic data. CRLC supported the analysis of oil content and obtained data from the device. MIZ is the creator and coordinator of the study, and provided financial support. EGOC wrote the manuscript. All authors read, edited, and approved the final manuscript.

Funding

This work was supported by the São Paulo Research Foundation, FAPESP (grant 2019/20307–0 and by the Thematic Project—grant 14/23591–7). The publication fee was supported by Luiz de Queiroz Agricultural Studies Foundation, Fealq.

Data availability

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1107811. The barcodes of the sequenced samples are available in the GitHub repository (https://github.com/evellyngocouto/Macauba_GWAS). The genotypic and phenotypic datasets generated and analysed during the current study are also there.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Genetics, "Luiz de Queiroz" College of Agriculture, São Paulo University, (ESALQ/USP), Piracicaba, Brazil. [2]Department of Plant Biology, University of Campinas (UNICAMP), Campinas, Brazil. [3]Department of Crop Soil, and Enviromental Sciences, Center of Agrcultural Data Analytics, University of Arkansas, Fayetteville, USA. [4]Research Center of Plant Genetic Resources, Campinas Agronomic Institute, Campinas, Brazil. [5]Polo Centro Sul, São Paulo Agency for Agribusiness Technology (APTA), Piracicaba, Brazil.

Received: 20 April 2024 Accepted: 11 November 2024
Published online: 26 November 2024

References

1. De Lima NE, Carvalho AA, Meerow AW, Manfrin MH. A review of the palm genus Acrocomia: Neotropical green gold. Org Divers Evol. 2018;18:151–61.
2. Scariot A, Lleras E, Hay JD. Flowering and Fruiting Phenologies of the Palm Acrocomia aculeata: Patterns and Consequences. Biotropica. 1995;27:168.
3. Lorenzi H, Noblick L, Kahn F, Ferreira E. Flora Brasileira: Arecaceae (palmeiras). Nova Odessa, SP: Instituto plantarum; 2010.
4. Colombo CA, Chorfi Berton LH, Diaz BG, Ferrari RA. Macauba: a promising tropical palm for the production of vegetable oil. OCL. 2018;25:D108.
5. Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL, et al. The search for loci under selection: trends, biases and progress. Mol Ecol. 2018;27:1342–56.
6. Colombo CA, Chorfi Berton LH, Diaz BG, Ferrari RA. Macauba: a promising tropical palm for the production of vegetable oil. OCL. 2018. https://doi.org/10.1051/ocl/2017038.
7. Vargas-Carpintero R, Hilger T, Mössinger J, Souza RF, Barroso Armas JC, Tiede K, et al. Acrocomia spp.: neglected crop, ballyhooed multipurpose palm or fit for the bioeconomy? A review. Agron Sustain Dev. 2021;41:75.
8. Vargas-Carpintero R, Hilger T, Tiede K, Callenius C, Mössinger J, Souza RF, et al. A Collaborative, Systems Approach for the Development of Biomass-Based Value Webs: The Case of the Acrocomia Palm. Land. 2022;11:1748.
9. Ciconini G, Favaro SP, Roscoe R, Miranda CHB, Tapeti CF, Miyahira MAM, et al. Biometry and oil contents of Acrocomia aculeata fruits from the Cerrados and Pantanal biomes in Mato Grosso do Sul. Brazil Ind Crops Prod. 2013;45:208–14.
10. Coimbra MC, Jorge N. Fatty acids and bioactive compounds of the pulps and kernels of Brazilian palm species, guariroba (*Syagrus oleraces*), jerivá (*Syagrus romanzoffiana*) and macaúba (*Acrocomia aculeata*). J Sci Food Agric. 2012;92:679–84.
11. Hiane PA, Filho MMR, Ramos MIL, Macedo ML. Bocaiuva, Acrocomia aculeata (Jacq.) Lodd., pulp and kernel oils: characterization and fatty acid composition. Braz J Food Technol. 2005;8:256–9.
12. Silva JC, Barrichelo LEG. Endocarpos de Macaúba e de Babaçu comparados a madeira de Eucaliptus grandis na produção de carvão vegetal. 1986;34:31–4.
13. Aires GCM, de Carvalho Junior RN. Potential of Supercritical Acrocomia aculeata Oil and Its Technology Trends. Appl Sci. 2023;13:8594.
14. Madeira DDC, Motoike SY, Simiqueli GF, Kuki KN, De Melo GS, Rigolon TCB, et al. Phenotypic characterization and genetic diversity of macauba (Acrocomia aculeata) accessions based on oil attributes and fruit biometrics. Genet Resour Crop Evol. 2024. https://doi.org/10.1007/s10722-024-01856-0.
15. Clement CR. 1492 and the loss of amazonian crop genetic resources. I. The relation between domestication and human population decline. Econ Bot. 1999;53:188–202.
16. Clement CR, Casas A, Parra-Rondinel FA, Levis C, Peroni N, Hanazaki N, et al. Disentangling Domestication from Food Production Systems in the Neotropics. Quat. 2021;4:4.
17. Abreu AG, Priolli RHG, Azevedo-Filho JA, Nucci SM, Zucchi MI, Coelho RM, et al. The genetic structure and mating system of Acrocomia aculeata (Arecaceae). Genet Mol Biol. 2012;35:116–21.
18. Cruz CD. Biometria aplicada ao estudo da diversidade genética/ Cosme Damião Cruz, Fábio Medeiros Ferreira, Luiz Alberto Pessoni. Visconde do Rop Branco: Suprema; 2011. p. 620.
19. Farias Neto JTD, Clement CR, Resende MDVD. Estimativas de parâmetros genéticos e ganho de seleção para produção de frutos em progênies de polinização aberta de pupunheira no estado do Pará. Brasil Bragantia. 2013;72:122–6.
20. Díaz BG, Zucchi MI, Alves-Pereira A, De Almeida CP, Moraes ACL, Vianna SA, et al. Genome-wide SNP analysis to assess the genetic population structure and diversity of Acrocomia species. PLoS ONE. 2021;16:e0241025.
21. Laviola BG, Dos Santos A, Rodrigues EV, Teodoro LPR, Teodoro PE, Rosado TB, et al. Structure and genetic diversity of macauba [Acrocomia aculeata (Jacq.) Lodd. ex Mart.] approached by SNP markers to assist breeding strategies. Genet Resour Crop Evol. 2022;69:1179–91.
22. Costa AM, Motoike SY, Corrêa TR, Silva TC, Coser SM, Resende MDVD, et al. Genetic parameters and selection of macaw palm (Acrocomia aculeata) accessions: an alternative crop for biofuels. Crop Breed Appl Biotechnol. 2018;18:259–66.
23. Coser SM, Motoike SY, Corrêa TR, Pires TP, Resende MDV. Breeding of Acrocomia aculeata using genetic diversity parameters and correlations to select accessions based on vegetative, phenological, and reproductive characteristics. Genet Mol Res. 2016;15:1–11.
24. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th edition. New York, NY: Longman Group Limited: Edinburgh; 1996.
25. Alves-Pereira A, Zucchi MI, Clement CR, Viana JPG, Pinheiro JB, Veasey EA, et al. Selective signatures and high genome-wide diversity in traditional Brazilian manioc (Manihot esculenta Crantz) varieties. Sci Rep. 2022;12:1268.
26. Morales-Marroquin JA, Diaz-Hernandez BG, Vianna SA, Alves-Pereira A, De Araujo-Batista CE, Colombo CA, et al. Genetic variations associated with adaptation processes in Acrocomia palms: A comparative study across the Neotropic for future crop improvement. bioRxiv. 2024;2024.08.15.608149. https://doi.org/10.1101/2024.08.15.608149.
27. Babu K, Mathur RK, Venu MVB, Shil S, Ravichandran G, Anita P, et al. Genome-wide association study (GWAS) of major QTLs for bunch and oil yield related traits in Elaeis guineensis L. Plant Sci. 2021;305:110810.
28. Babu BK, Mathur RK, Ravichandran G, Anita P, Venu MVB. Genome wide association study (GWAS) and identification of candidate genes for yield and oil yield related traits in oil palm (Eleaeis guineensis) using SNPs by genotyping-based sequencing. Genomics. 2020;112:1011–20.
29. Osorio-Guarín JA, Garzón-Martínez GA, Delgadillo-Duran P, Bastidas S, Moreno LP, Enciso-Rodríguez FE, et al. Genome-wide association study (GWAS) for morphological and yield-related traits in an oil palm hybrid (Elaeis oleifera x Elaeis guineensis) population. BMC Plant Biol. 2019;19:1–11.
30. Wibowo CS, Apriyanto A, Ernawan R, Neing D, Susilo R, Cordell HJ, et al. Genetic variants associated with leaf spot disease resistance in oil palm (Elaeis guineensis): A genome-wide association study. Plant Pathol. 2023;72:1626–36.
31. Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, et al. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics. 2015;105:288–95.
32. Seng T-Y, Ritter E, Mohamed Saad SH, Leao L-J, Harminder Singh RS, Qamaruz Zaman F, et al. QTLs for oil yield components in an elite oil palm (Elaeis guineensis) cross. Euphytica. 2016;212:399–425.
33. Malosetti M, Ribaut JM, Vargas M, Crossa J, Van Eeuwijk FA. A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.). Euphytica. 2008;161:241–57.
34. François O, Caye K. Naturalgwas: An R package for evaluating genomewide association methods with empirical data. Mol Ecol Resour. 2018;18:789–97.
35. Voichek Y, Weigel D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet. 2020;52:534–40.
36. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11:407–9.

37. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016;32:1981–9.
38. Joo JWJ, Kang EY, Org E, Furlotte N, Parks B, Hormozdiari F, et al. Efficient and Accurate Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure. Genetics. 2016;204:1379–90.
39. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–8.
40. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genet. 2016;12.
41. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience. 2019;8(2):giy154.
42. Fernandes SB, Zhang KS, Jamann TM, Lipka AE. How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy? Front Genet. 2021;11:602526.
43. Furlotte NA, Eskin E. Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model. Genetics. 2015;200:59–68.
44. Pritikin JN, Neale MC, Prom-Wormley EC, Clark SL, Verhulst B. GW-SEM 2.0: Efficient, Flexible, and Accessible Multivariate GWAS. Behav Genet. 2021;51:343–57.
45. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
46. Fernandes SB, Casstevens TM, Bradbury PJ, Lipka AE. A multi-trait multi-locus stepwise approach for conducting GWAS on correlated traits. Plant Genome. 2022;15.
47. Manfio CE, Motoike SY, Santos CEMD, Pimentel LD, Queiroz VD, Sato AY. Repetibilidade em características biométricas do fruto de macaúba. Ciênc Rural. 2011;41:70–6.
48. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;1:1–48.
49. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. URL https://www.R-project.org/.
50. Covarrubias-Pazaran G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLoS ONE. 2016;11.
51. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE. 2012;7.
52. Doyle JJ, Doyle JL. Isolation ofplant DNA from fresh tissue. Focus. 1990;13:39–40.
53. Poland JA, Rife TW. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome. 2012;5:plantgenome2012.05.0005.
54. Andrews, S. FASTQC. A quality control tool for high throughput sequence data. 2010.
55. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. *Stacks* : Building and Genotyping Loci *De Novo* From Short-Read Sequences. G3 GenesGenomesGenetics. 2011;1:171–82.
56. Singh R, Ong-Abdullah M, Low E-TL, Manaf MAA, Rosli R, Nookiah R, et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature. 2013;500:335–9.
57. Bazzo BR, De Carvalho LM, Carazzolle MF, Pereira GAG, Colombo CA. Development of novel EST-SSR markers in the macaúba palm (Acrocomia aculeata) using transcriptome sequencing and cross-species transferability in Arecaceae species. BMC Plant Biol. 2018;18:276.
58. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
60. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012.
61. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
62. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021;108:1880–90.
63. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103:338–48.
64. Goudet J. HIERFSTAT, a package for R to compute and test hierarchical *F* -statistics. Mol Ecol Notes. 2005;5:184–6.
65. Jombart T. *adegenet*: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24:1403–5.
66. Wang J, Zhang Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics Bioinformatics. 2021;19:629–40.
67. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
68. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44:825–30.
69. Peterson RA, Cavanaugh JE. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. J Appl Stat. 2020;47:2312–27.
70. Filzmoser P, Gschwandtner M. Package 'mvoutlier.' 2018.
71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
72. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
73. Heberle H, Meirelles GV, Da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics. 2015;16:169.
74. Teixeira LC. Potencialidades de oleaginosas para produção de biodiesel. Informe agropecuário. 2005;18–27.
75. Mazzottini-dos-Santos HC, Ribeiro LM, Mercadante-Simões MO, et al. Floral structure in Acrocomia aculeata (Arecaceae): evolutionary and ecological aspects. Plant Syst Evol. 2015;301:1425–40.
76. Montoya SG, Motoike SY, Kuki KN, Couto AD. Fruit development, growth, and stored reserves in macauba palm (Acrocomia aculeata), an alternative bioenergy crop. Planta. 2016;244:927–38.
77. Berton, Luiz Henrique Chorfi. Avaliação de populações naturais, estimativas de parâmetros genéticos e seleção de genótipos elite de macaúba (Acrocomia aculeata). Instituto Agronômico de Campinas; 2013.
78. Domiciano Silva Rosado R, Barbosa Rosado T, Damião Cruz C, Gomes Ferraz A, Haa Carson Schwartzhaupt Da Conceição LD, Galveas Laviola B. Genetic parameters and simultaneous selection for adaptability and stability of macaw palm. Sci Hortic. 2019;248:291–6.
79. Fernandes SB, Dias KOG, Ferreira DF, Brown PJ. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theor Appl Genet. 2018;131:747–55.
80. Domiciano GP, Alves AA, Laviola BG, Conceição LDHCSD. Parâmetros genéticos e diversidade em progênies de Macaúba com base em características morfológicas e fisiológicas. Ciênc Rural. 2015;45:1599–605.
81. Regazzi AJ. Modelos biométricos aplicados ao melhoramento genético. Editora UFV; 2011.
82. John Dransfield, Natalie W. Uhl, Conny B. Asmussen, William J. Baker, Madeline M. Harley, Carl E. Lewis. Genera Palmarum: The Evolution and Classification of Palms. International Palm Society; 2014.
83. Rafaël Govaerts, John Dransfield. World Checklist of Palms. Royal Botanic Garden, Kew; 2005.
84. Francisconi AF, Marroquín JAM, Cauz-Santos LA, Van Den Berg C, Martins KKM, Costa MF, et al. Complete chloroplast genomes of six neotropical palm species, structural comparison, and evolutionary dynamic patterns. Sci Rep. 2023;13:20635.
85. Stich B, Melchinger AE. An introduction to association mapping in plants. CABI Rev. 2010;:1–9.
86. Zhang H, Ma X, Shi T, Song Q, Zhao H, Ma D. NSA2, a novel nucleolus protein regulates cell proliferation and cell cycle. Biochem Biophys Res Commun. 2010;391:651–8.
87. Ciftci-Yilmaz S, Mittler R. The zinc finger network of plants. Cell Mol Life Sci. 2008;65:1150–60.
88. Naested H, Holm A, Jenkins T, Nielsen HB, Harris CA, Beale MH, et al. *Arabidopsis VARIEGATED 3* encodes a chloroplast-targeted, zinc-finger

protein required for chloroplast and palisade cell development. J Cell Sci. 2004;117:4807–18.
89. Ding L-N, Li M, Wang W-J, Cao J, Wang Z, Zhu K-M, et al. Advances in plant GDSL lipases: from sequences to functional mechanisms. Acta Physiol Plant. 2019;41:151.
90. Cao Y, Han Y, Meng D, Abdullah M, Yu J, Li D, et al. Expansion and evolutionary patterns of GDSL-type esterases/lipases in Rosaceae genomes. Funct Integr Genomics. 2018;18:673–84.
91. Hartl FU. Molecular chaperones in cellular protein folding. Nature. 1996;381:571–80.
92. Hartl FU. Chaperone-assisted protein folding: the path to discovery from a personal perspective. Nat Med. 2011;17:1206–10.
93. Sharma SK, De Los RP, Christen P, Lustig A, Goloubinoff P. The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase. Nat Chem Biol. 2010;6:914–20.
94. Hassan H, Amiruddin MD, Weckwerth W, Ramli US. Deciphering key proteins of oil palm (*Elaeis guineensis* Jacq.) fruit mesocarp development by proteomics and chemometrics. Electrophoresis. 2019;40:254–65.
95. Lee C, Teng Q, Zhong R, Ye Z-H. The Four Arabidopsis REDUCED WALL ACETYLATION Genes are Expressed in Secondary Wall-Containing Cells and Required for the Acetylation of Xylan. Plant Cell Physiol. 2011;52:1289–301.
96. Manabe Y, Nafisi M, Verhertbruggen Y, Orfila C, Gille S, Rautengarten C, et al. Loss-of-Function Mutation of *REDUCED WALL ACETYLATION2* in Arabidopsis Leads to Reduced Cell Wall Acetylation and Increased Resistance to *Botrytis cinerea*. Plant Physiol. 2011;155:1068–78.
97. Nafisi M, Stranne M, Fimognari L, Atwell S, Martens HJ, Pedas PR, et al. Acetylation of cell wall is required for structural integrity of the leaf surface and exerts a global impact on plant stress responses. Front Plant Sci. 2015;6.
98. Zhang K, Yao D, Chen Y, Wen H, Pan J, Xiao T, et al. Mapping and identification of CsSF4, a gene encoding a UDP-N-acetyl glucosamine-peptide N-acetylglucosaminyltransferase required for fruit elongation in cucumber (Cucumis sativus L.). Theor Appl Genet. 2023;136:54.

Publisher's Note