Genome-Wide Association Insights into the Genomic Regions Controlling Oil Production Traits in *Acrocomia aculeata* (neotropical native palm)

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

[1] Agência Paulista de Tecnologia dos Agronegócios (APTA), pólo de Piracicaba, Brazil
[2] Universidade de Campinas, Campinas, Brazil
[3] University of Arkansas, Arkansas, United States
[4] Instituto Agronômico de Campinas, Campinas, Brazil

* Correspondent author: mizucchi@sp.gov.br, evellyncoutoo@gmail.com

ABSTRACT

Macaúba (*Acrocomia aculeata*) is a non-domesticated neotropical palm that has been attracting attention for economical use due to its great potential for oil production, comparable to the commercially used oil palm (*Elaeis guineenses*). This palm has an annual production, and its fruit harvest occurs without the destruction of the plant, allowing for sustainable management. Besides, it can be used in integrated Crop-Livestock-Forestry systems, preventing devastating effects on the environment. With genetic improvement, oil and biofuel production can increase and make this species more competitive. However, that requires an understanding of the architecture of the agronomic traits. For instance, 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 the macaúba domestication. In this context, single-trait and multi-trait GWAS models were performed to identify candidate genes related to oil production traits in macaúba. We randomly selected 201 palms from a natural population and collected phenotypic data in two years of production. Phenotypic measurements of 13 traits involved with fruit production, processing, and pulp oil content were obtained, and data analysis was performed using a mixed linear model. Genotyping was performed with SNP markers, following the protocol of genotyping-by-sequencing. The SNP calling was performed using three different strategies since macaúba does not have a reference genome: using i) de novo sequencing, ii) the *Elaeis guineenses* Jacq. reference genome and iii) the macaúba transcriptome sequences. The quality control of the SNP data was performed considering minor allele frequency ≥ 0.01, a maximum of 30% of missing data per locus, and 45% of missing data per individual. Missing data was inputted with the software Beagle 5.3. Single-trait analysis was fitted using five different models from GAPIT in the software R, while multi-trait analysis was fitted using a multivariate stepwise method implemented in the software

TASSEL. Multi-trait analyses were conducted in all pairwise trait combinations. Results showed statistically significant differences in all phenotypic traits studied, and heritability values ranged from 63 to 95%. Genetic correlation between traits ranged from -0.47 to 0.99. Gene annotation in Blast2Go detected 15 candidate genes in seven phenotypic traits in the single-trait GWAS, ten in the oil palm genotypic dataset, four in the de novo dataset, and one in the transcriptome dataset, in the multi-trait GWAS, four candidate genes were detected in ten traits combination. Gene mapping for the candidate genes identified similarities with *Elaeis guineensis* and *Phoenix dactylifera*. Candidate genes are responsible for metal ion binding and transport, protein transportation, DNA repair, and other cell regulation biological process. We provide new insights on genomic regions that mapped candidate genes involved in macaúba oil production phenotypes. These potential candidate genes need to be confirmed for future targeted functional analyses and multi-trait associations need to be scrutinized to investigate the presence of pleiotropic or linked genes. Associated markers to the traits of interest may be valuable resources for the development of marker-assisted selection in macaúba for its domestication and pre-breeding.

Keywords: Neotropical Oil Palm, macaúba, coyol, macaw palm, associative mapping, multi-trait associations, oil content, biofuel.

1 INTRODUCTION

Facing the need to implement renewable energy sources to reduce environmental impacts, biofuels have emerged as a promising alternative for reducing carbon emissions from non-renewable fossil sources. Biofuels are derived from biomass and can replace, partially or completely, fuels derived from petroleum and natural gas in combustion engines or in other types of energy generation. They are alternative energy sources that have a low pollutant emission rate compared to fóssil fues. The main raw material for biodiesel production comes from vegetable oil that can be extracted from plants, such as soybeans, maize, sugar cane, oil palm, sunflower, among many other species, of which the palm macaúba (*Acrocomia aculeata* (Jacq.) Lood. ex Mart) have stood out, due to its high fruit production and sizable oil content in the fruit mesocarp (Colombo et al., 2018).

Macaúba or Coyol is a neotropical palm (Bactridinae), its distribution goes from Mexico to the Northen region of Argentina. In Brazil, macaúba occurs in the states of Ceará, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Tocantis, Roraima, Amazonas and São Paulo (Scariot et al., 1995, Lima et al., 2018; Lorenzi et al., 2010).

The genus *Acrocomia* presents great potential to become commercially exploited, being *Acrocomia aculeata* the most targeted species because of the high oleic

concentration in its fruits. This palm has an annual production, allowing sustainable management due to its adaptation for growing on dry biomes, bare ground, rangeland, or anthropogenic landscapes. Macaúba starts producing fruit around the fifth year of life, and guarantees production for up to 50 years, avoiding excessive soil management (Teixeira, 2005); and/or degradation of vulnerable humid biomes, like Tropical Rain Forests. Other oilseeds need precipitation rates from ~2000 millimeters each year for their proper cultivation. Thus, macaúba can be used in integrated Crop-Livestock-Forest systems, or in Regenerative Land Management, optimizing its use and production gains.

In industry, all macaúba fruits parts (epicarp, mesorcarp, endocarp and endosperm) can be used for raw material production. Fruits present oil in the mesocarp and endosperm. The oil content in the mesocarp can achieve 70%, rich in oleic acid (Ciconini et al., 2013) targeted in the biofuel industry; while the oil content in the endosperm is higher than 50%, rich in lauric acid used in the cosmetics industries and other saponification products (Coimbra, 2012). Furthermore, the fruits are abundant in amino acids, carbohydrates, and minerals (Hiane et al., 2005), used in the animal and human food industry. Also, the endocarp has a high lignin content that could be used in the coal production (Silva et al., 1986).

Despite the extraction of macaúba raw material have many uses and high potential for biodiesel production, this palm is a incipiently domesticated species, and most of oil consumed is obtained from extractivism by family farming (Abreu et al., 2012). Diaz et al. (2021) studied genetic diversity in *Acrocomia ssp*. distributed in Brazil, and found high values of genetic variability, so higher productive native palms can be selected from natural populations to start the first pre-breeding cycles. The selection of parental plants with potential for productivity in a natural population is based on their phenotypic traits; and the correlation estimates between its evaluated traits enables the selection and understanding of their genetic basis. The correlation depends on the association of the existing variability between the selected traits, such as phenotypic, environmental and genetic, whose latter is the most importante aspect because indicates high heritability between traits, a measurement of great interest in plant breeding. Previous studies in macaúba have observed high genotypic correlations between the traits of plant height and the number of green and dry leaves, stem length, stipe thickness, and thorn occurrence; as well as between leaf length and stem length, leaf length and stipe thickness, among other traits (Berton, 2013). High values of genetic correlations between traits of interest can facilitate the breeding cycle, once breeders can focus in the choice of few traits.

Furthermore, the existence of a high correlation between traits indicates that there are pleiotropic and/or linked genes in the control of both traits evaluated (Falconer; Mackay, 1996; Cruz; Regazzi, 2004). Thus, if supposedly we consider that three traits (A, B and C) are controlled by the same gene, the improvement of the trait A will also correspond to the improvement of the other two, B and C. That is, the breeder's efforts in selecting individuals whose trait A is more promising, indicates that he will also be selecting the best trait B and C at the same time, saving time.

In plant breeding, interesting agronomic traits are controlled mostly by several genes qualified as quantitative traits. Quantitative Trait Loci (QTLs) are widely used to name chromosomal regions that have loci controling the expression of the quantitative traits (Falconer; Mackay, 1996; Mackay, 2001). Recently, molecular markers are used for mapping QTLs for several traits of interest, allowing a better understanding of the genetic basis of the correlation between traits through the study of linked and pleiotropy genes (Falconer; Mackay, 1996; Malosetti et al. al., 2008). This can be directly applied to genetic improvement programs. QTL mapping studies are based on association analysis, also known as genome-wide association studies (GWAS) (Oraguize et al., 2007). GWAS attempt to evaluate causal relationships between genetic variants and a phenotype of interest by surveying a genome-wide set of genetic polymorphisms in a large number of individuals (François & Caye, 2018). The focus of this methodology is to find old mutations that have been maintained throughout the evolution of the population (Zhu et al., 2008), so that the association detected is attributed to linkage disequilibrium between markers and functional polymorphisms in a set of evaluated genotypes (Yu et al., 2006). The GWAS approach was possible due to the current availability of SNP markers, based on next generation sequencing (NGS), which allowed whole genome screening (Poland & Rife, 2012). Decreasing genotyping costs make GWAS an attractive option for facilitating candidate gene identification.

Single-trait and multi-trait GWAS approaches are currently employed to detect cross-phenotype associations (Zhou and Stephens, 2014; Cichonska et al., 2016; Joo et al., 2016). Some of the first GWAS models, implemented to analyse phenotype-genotype association, such as general linear models and mixed-linear models (MLMs), were single-locus, single-trait GWAS, created to implement covariates along with kinship matrices (Yu et al., 2006). However, these simple models resulted in false negatives caused by weakened associations, due to population structure. In this case, the models CMLMs (compressed MLMs) (Zhang et al., 2010) and SUPER (Wang et al., 2014) were

developed. Additionally, to evaluate big datasets, while also reducing false positives and negatives, multi-locus GWAS, such as FarmCPU (Liu et al. 2016) and BLINK (Huang et al., 2019), were developed. Unlike the single-trait models, multi-trait models enable the quantification of simultaneous loci contributions for multiple traits in GWAS studies (Zhou & Stephens, 2014). A lot off software packages have implemented multi-trait GWAS models (Furlotte & Eskin, 2015; Pritikin et al., 2021; Zhou & Stephens, 2014); and recently a new model was implemented in the software TASSEL (Bradbury et al., 2007) by Fernandes et al. (2021). This model, named multi-trait multi-locus stepwise (MSTEP) showed to be more promising in a scenario where multiple traits are controlled by both pleiotropic and nonpleiotropic causal mutations (Fernandes et al., 2021). The authors even concluded that the multi-trait models outperformed the univariate multi-locus model, especially when analyzing a trait of low heritability. In macaúba, genomic studies using SNP markers were firstly applied by Diaz et al. (2021). However, until now, there are none studies applying GWAS for *Acrocomia aculeata*, and results related to the genomic association with the traits responsible for oil production are of great interest for understanding the genetic architecture in this specie, even to facilitate its domestication and pre-breeding. In addition, markers detected in the multi-trait GWAS can guide studies about the detection of linked or pleiotropic genes, which would enrich the genetic information about this palm. So, here we present the first study using a GWAS approach for understanding the genomic regions controlling oil production traits in *Acrocomia aculeata*.

2 MATERIAL AND METHODS

2.1 Plant materials and phenotyping

We studied a panel comprise of 201 *Acrocomia aculeata* palms from a natural population located in Dourado, São Paulo state, Brazil (geographical coordinates 22° 6' 13" S, 48° 18' 50" W). Palms were randomly selected in four subpopulations located in four rural areas (**Figure 1**), and data was collected through two different seasons of fruit production in the years of 2019/2020 and 2020/2021. Subpopulation 1 is formed by individuals numbered 1 to 100, subpopulation 2 by individuals numbered 101 to 130, subpopulation 3 by individuals numbered 131 to 184 and subpopulation 4 by individuals numbered 185 to 201.

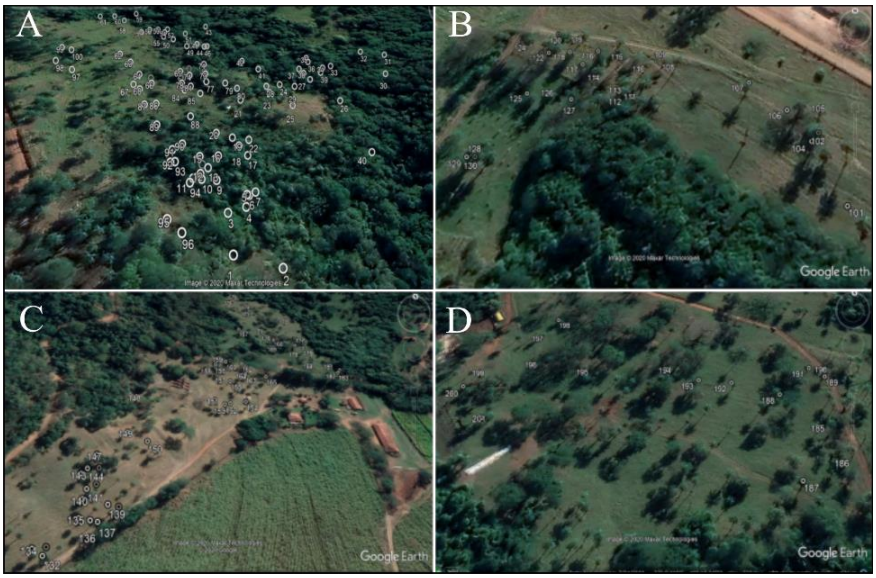

Figure 1. Rural area where *Acrocomia aculeata* palms were collected. A) Subpopulation 1. B) Subpopulation 2. C) Subpopulation 3. D) Subpopulation 4.

Different phenotypic measurements were obtained from the palms: height (H - in meters), stipe (trunk) diameter at breast height (DHB - in centimeters), leaves number (LN), leaves length (LL - in centimeters), leaves needles number (NN), leaves needles length (NL - in centimeters), leaves needles width (NW - in centimeters) and total fruit weight (FW – in grams). The fruit characterization was based on 6 fruits from each palm, in three repetitions, being that two fruits comprised one repetition. After obtaining the total weight of the fruits, they were manually separated into four fractions: (1) epicarp (husk), (2) mesorcarp (pulp), (3) endocarp (shell), and (4) endosperm (almond) for fresh and dry mass (**Figure 2B**). The fresh mass of the fruit fractions was dried in a ventilated oven at 36ºC for 36 hours. We identify these traits as: husk fresh mass (HFM), pulp fresh mass (PFM), endocarp fresh mass (EFM), almond fresh mass (AFM), husk dry mass (HDM), pulp dry mass (PDM), endocarp dry mass (EDM) and almond dry mass (ADM). The weights of the four dry fractions were added to obtain the value of the total weight of the dry fruit (FDW).

2.2 Oil content quantification by NIRS

Oil content (OC) was quantified from dry mesocarp mass (**Figure 2C**) using Near-infrared Spectroscopy (NIRS), values are expressed in percentage (% or g/100g). The dry mesocarps of the three replicates were crushed in the analytical mil IKA–A11. Diffuse reflectance of the mesocarp samples was measured in rectangular cells (48 x 58 mm), using a FOSS NIRSystems 6500 spectrophotometer. The spectra measurements of the dry

mesocarp mass 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) for predicting the oil content using the NIRs spectros graphs.

Figure 2. Plant material and phenotyping. A) *Acrocomia aculeata* in the field. B) Fruit fractions in three repetitions; considering from left to righ: epicarp (husk), mesocarp (pulp), endocarp, almond. C) Diferent shades of dry mesocarp flours used in oil content quantification.

2.3 Genotypic dataset

Genotyping was performed with markers based on single-nucleotide polymorphism (SNPs), following the protocol of genotyping-by-sequencing using two restriction enzymes (ddGBS) (Poland et al., 2012). Genotypic data material was collected from the 201 *Acrocomia aculeata* palms selected in this study. The genomic DNA was isolated from leaf material using the Doyle & Doyle (1990) protocol. 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 was 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 30ng.µl-1.

2.4 GBS library preparation and high-throughput sequencing

For obtaining the SNPs, we prepared three 96-plex genomic libraries using the ddGBS technique according to the protocol described by Poland et al. 2012 with modifications, the same used by Díaz et. al., 2021. 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 (BioRad) 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 sample libraries prepared were sequenced on a single run, in an Illumina HiSeq3000, with single-end and 101bp configurations. The overall quality of the sequencing of GBS libraries was evaluated with the FastQC program (ANDRWES, 2010). Quality control and *demultiplex* were performed with the *process_radtags* module of the Stacks 1.42 program (Catchen et al., 2011), where low-quality reads were removed.

2.5 SNP calling

The SNP calling was performed using three different strategies because there is no reference genome available for *Acrocomia aculeata*. First, (i) using the *de novo* pipeline (Stacks v.1.42) (Catchen et al., 2011) based on the alignment of the reads obtained in the genotyping, (ii) using the genome of *Elaeis guineensis* var *tenera* (oil palm) as reference and (iii) using the transcriptome of *Acrocomia aculeata* (Bazzo et al., 2018).

SNP calling using de novo pipeline

The identification of SNP was carried out using the *de novo* pipeline in the software Stacks *v*. 1.42 (Catchen et al., 2011). After the quality control, samples were demultiplexed using process-rad-tags in Stacks. Illumina reads were trimmed to 90 bp, 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). Subsequently, a locus catalog was built using the *cstacks* module, allowing a maximum of two differences between stacks

(BOUTET et al., 2016) from different individuals. 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 per sequence.

SNP calling using reference sequences

The *bwa-mem* algorithm of the *bwa* 0.7.17 program (Li, 2013) was used to align the sequences of each sample to the genome of *Elaeis guineensis* var *tenera* (oil palm) EG5 (NCBI GCA_000442705.1) and to the transcriptome sequencing of *Acrocomia aculeata* (Bazzo et al., 2018). Alignment files were processed with SAMtools (Li et al., 2009) and Picard programs (http://broadinstitute.github.io/picard). SNP identification was performed using the program *freebayes* 1.3.4 (Garrison & Marth, 2012) with the configuration --standard_filters. VCFtools 0.1.17 (Danecek et al., 2011) and bcftools 0.1.12 (Li et al., 2009) programs were used to filter SNP markers, retaining all SNP per sequence.

2.6 Quality control, imputation and organization of genotypic data

The following criteria were used for filtering SNPs in the three SNP calling strategies mentioned above: 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. After filtering, we identified a total of 27410 SNPs in 153 individuals for the *de novo* pipeline, 10444 SNPs in 158 individuals using the oil palm reference genome, and 4329 SNPs in 167 individuals using the transcriptome data. Missing data from the three datasets used in this study was input into the Beagle 5.3 software (Browning et al., 2018, 2021).

Chromosome information was added to the three genotypic datasets to perform GWAS analyses. For *de novo* and *transcriptome* dataset, which do not have a reference genome, we considered that all the SNP markers belonged to the same chromosome. For the oil palm dataset, chromosome information was obtained from the NCBI GCA_000442705.1. The oil palm reference genome holds information from 16 chromosomes and, also, from the complete chloroplast genome (reference sequence named NC_017602.1). Thus, the sequences NC_017602.1 obtained in our dataset, which referred to the chloroplast genome, were excluded. We also obtained sequences that have not yet been allocated in the oil palm genome, named "SNW". In this case, these sequences were allocated by us to a chromosome that does not exist, number 17, to

promote the comprehension of the results.

2.6 Phenotypic data analysis

Statistical analysis of the phenotypic data was performed using a mixed linear model in the lme4 package in the R software, version 4.0.5 (R Core Team, 2021). The year information was used as a repetition because our data is from natural *Acrocomia aculeata* populations, which means that this study did not have an experimental design. Statistical analysis of fruit characteristics was performed considering the mean of the three replications. Outliers were detected and removed by using Shapiro-Wilk test at an overall significance level of 5%. The statistical model used to study the significance of phenotypic characteristics was:

$$Y_{ij} = \mu + g_i + y_j + \varepsilon_{ij}$$

where Y_{ij} is the value obtained from each character under study, μ is the general mean, g_i is the random effect of the germplasm source i, y_j is the fixed effect of years j and ε_{ij} is the residual.

The significance of random effect was estimated by Likelihood Ratio Test at 5% probability and the variance components were used to estimate the random sense heritability (H^2):

$$H^2 = \frac{\sigma^2{}_G}{\sigma^2{}_G + \frac{\sigma^2{}_e}{r}}$$

where $\sigma^2{}_G$ is the genotypic variance, $\sigma^2{}_e$ is the residual variance and r is the repetition.

Pearson's phenotypic correlation coefficients were calculated in all the characters studied, while genotypic correlation was calculated by multivariate linear mixed model by Likelihood methods (REML) using sommer R package, version 4.1.5. (Covarrubias-Pazaran, 2016).

2.7 Genetic diversity and population structure

Genetic diversity and population structure were analyzed individually for each genotypic dataset in two different R packages: hierfstat (Goudet, et al., 2005) and

adegenet (Jombart, 2008), respectively. In this analysis, the four *Acrocomia aculeata* subpopulations were considered. Genetic diversity was investigated according to the value of the total number of alleles, the observed and expected heterozygosity, and the inbreeding coefficient, while population structure was inferred by principal component analysis (PCA).

2.8 *Single-trait and Multi-trait* GWAS

To conduct single-trait GWAS, five different statistical models ranging from single-locus to multilocus were compared using all the phenotypic characters studied, as well as the three genotypic datasets. The five statistical models from the GAPIT package (version 3) (Wang et al., 2021) included: (i) general linear model (GLM) (Price et al., 2006), (ii) MLM (Yu et al., 2006), (iii) multiple loci MLM (MLMM) (Segura et al., 2012), (iv) fixed and random model circulating probability unification (FarmCPU) (Liu et al., 2016), and (v) bayesian-information and linkage-disequilibrium iteratively nested keyway (BLINK) (Huang et al., 2019). To account for population structure, we utilize the "Model.selection = TRUE" and the "PCA.total=5", which selects the best number of principal components from 0 to 5. The relationship matrix was also calculated by GAPIT utilizing its default parameters.

The multi-trait analysis was fitted using a multivariate stepwise method (MSTEP) implemented in the software TASSEL, being conducted in all pairwise trait combinations (Fernandes et al., 2022). Prior to running MSTEP, each trait was normal quantile transformed using the orderNorm function from the package bestNormalize (Peterson, 2020). Next, multivariate outliers were removed based on the aq.plot function from the package mvoutlier (Filzmoser and Gschwandtner, 2021). 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.

2.9 Identification of candidate genes

Tag sequences containing the significative SNPs from the *de novo* dataset were manually recovered from the catalog generated by the *cstacks* module in the Stacks program. The sequences of the significative 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

significative SNPs in predicted gene regions for *Elaeis guineenses* (https://ftp.ncbi.nlm.nih.gov/genomes/all/ GCF/000/442/705/GCF_000442705.1_EG5/GCF_000442705.1_EG5_genomic.gff.gz) was verified with the *intersect* function from BEDTools v.2.30 program (Quinlan & Hall 2010), while the respective predicted protein sequences (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/442/705/ GCF_000442705.1_EG5/ GCF_000442705.1_EG5_protein.faa.gz) were recovered with the perl fasta_FetchSeqs.pl script. Blast2GO software (Conesa et al., 2005) was used to search for similarity between the sequences of significant SNPs and candidate genes. The *blastx* function was used for the *de novo* dataset and the *blastp* for the transcriptome and oil palm dataset. In the oil palm significant SNPs was detected within gene regions because the *Elaeis guineense* reference genome was used. For the other two genotypic datasets, which did not employ a genomic reference, candidate genes were searched for the similarity between sequences of significant SNPs and genes sequences from other species.

3. RESULTS

3.1 Phenotypic analysis

Mean phenotypic values and their range, estimates of genetic, phenotypic, and residual variances, coefficient of variation (CV), heritability (h^2) and likelihood ratio test (LRT) are showed in Table 1. The likelihood ratio test, performed for each of the phenotypic traits, detected significant differences among genotypes for all of them ($p < 0.001$) (**Table 1**). The ranges of the mean values within each trait studied showed a large interval. For OC, for example, the mean was 36.58%, ranging from 17.47% to 60.80%, while for FW, the mean was 76.51g, ranging from 39.90g to 147.53g. FW had the higher genetic, phenotypic and residual variance values (176.20g, 248.20g and 72.00g, respectively) while LL had the lower values (0.05cm, 0.07cm and 0.02cm, respectively). The value of the CV ranged from 22.38% in LN to 5.75% in LL. H, LL, OC, FW, PFM, EFM, HDM, EDM and ADM traits showed heritability values higher than 80%.

Table 1. Means (range), estimates of genetic, phenotypic, and residual variances ($\sigma^2{}_G$, $\sigma^2{}_p$, $\sigma^2{}_e$, respectively), coefficient of variation (CV%), heritability (h^2) and likelihood ratio test (LRT) for genotypes in eighteen evaluated traits.

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

H: height, DBH: diameter of the stem at breast height, LN: leaves number, LL: leaves length, NN: needles number, NL: needles length, NW: needles width, OC: oil content, FW: fruit weight, HFM: husk fresh mass, PFM: pulp fresh mass, EFM: endocarp fresh mass, AFM: almond fresh mass, FDW: fruit dry weight, HDM: husk dry mass, PDM: pulp dry mass, EDM: endocarp dry mass, ADM: almond dry mass.
*Significant according to the $\chi 2$ test ($\alpha = 0.01$).

Pairwise phenotypic and genetic correlation between the eighteen traits studied are showed in **Figure 3**. Phenotypic and genetic correlation values ranged among -1 to 1, in which -1 represent weak correlation between traits and 1, strong correlation. Results showed strong correlation between fruits traits in both analysis. However, phenotypic traits related to other features presented a weak correlation among them. The traits FW and FDW (0.93, 0.94), HFM and HDM (0.83, 0.90) PFM and PDM (0.91, 0.91), EFM and EDM (0.97, 1.00), AFM and ADM (0.93, 0.99) had strong phenotypic and genetic correlation, respectively, because they are the same traits. Nevertheless, the difference between them is the presence of humidity. Traits that presented the highest phenotypic and genetic correlations were: FW and HFM (0.78, 0.83); FW and PFM (0.94, 0.95), FDW and PFM (0.87, 0.88), FW and PDM (0.85, 0.86), FDW and HDM (0.81, 0.81), FDW and PDM (0.92, 0.92). For OC and NN, null and negative values were observed for

all the pairwise traits tested in both correlations. For example, NN and LL showed -0.30 for phenotypic correlation and -0.39 values for genetic correlation. The same was observed for NL and NN (-0.34, -0.47), respectively. OC had a phenotypic correlation varying from -0.00 to 0.12 between the pairwise evaluated traits while for genetic correlation, values ranged from 0.00 to -0.42 (**Figure 3**).

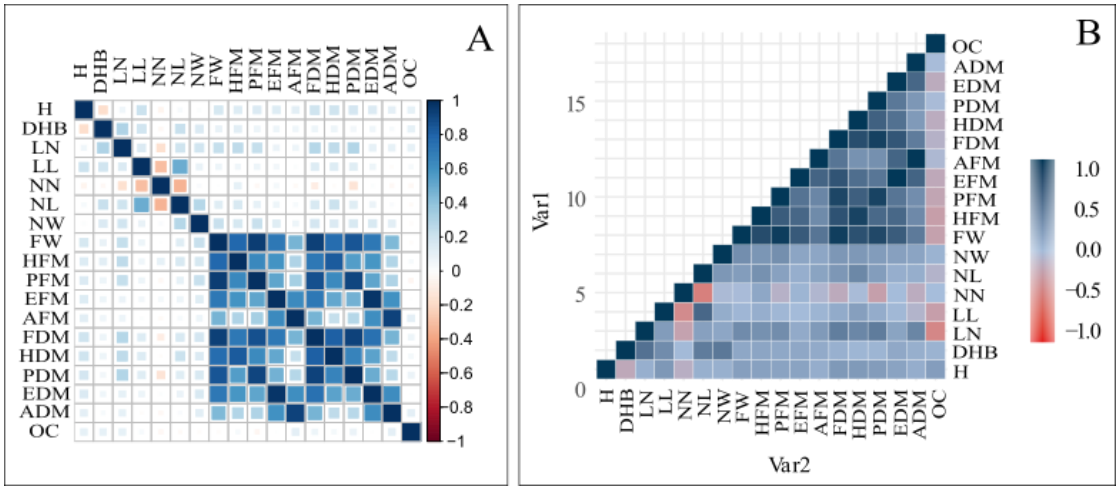

Figure 3. Correlation analysis among *Acrocomia aculeata* traits. A) Phenotypic correlation. B) Genetic correlation.

3.2 Genetic diversity, population structure and linkage disequilibrium

Genetic parameters values from the *Acrocomia aculeata* population studied were different for each SNP calling strategies (**Table 2**). The three genotypic datasets used in this study showed high values of heterozygosity and low values of inbreeding coefficient. *De novo* dataset showed the lowest diversity values, 0.15 to observed heterozygosity (H_o), 0.17 to expected heterozygosity (H_s), 0.18 to overall gene diversity (H_t) and 0.15 to Wright's inbreeding coefficient (*Fis*). Oil palm dataset showed the values of 0.41 to H_o, 0.30 to H_s, 0.31 to H_t and -0.35 to *Fis*, while transcriptome dataset showed the highest diversity values, 0.53 to H_o, 0.34 to H_s, 0.34 to H_t and -0.53 to *Fis*.

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

Genotypic dataset	Number of individuals	Number of SNP	H_o	H_s	H_t	Fis
De novo	153	27410	0.15	0.17	0.18	0.15
Oil palm	158	10444	0.41	0.30	0.31	-0.35
Transcriptome	167	4329	0.53	0.34	0.34	-0.53

Principal component analysis (PCA) in the three genotypic datasets revealed that individuals from subgroup 1 are slightly separated from individuals from subgroup 2, 3, and 4 (**Figure 3**). The four subgroups in the *de novo* dataset are evenly distributed in the graph, with some individuals of the subgroup 1 separated from the other subgroups. The same situation was observed in the transcriptome dataset, for the individuals from subgroup 1 and subgroups 2, 3, and 4. On the contrary, in the oil palm dataset, individuals from subgroups 2, 3, and 4 are tightly clustered, while the individuals from subgroup 1 are distant from the individuals of the other subgroups, with little overlap between them.

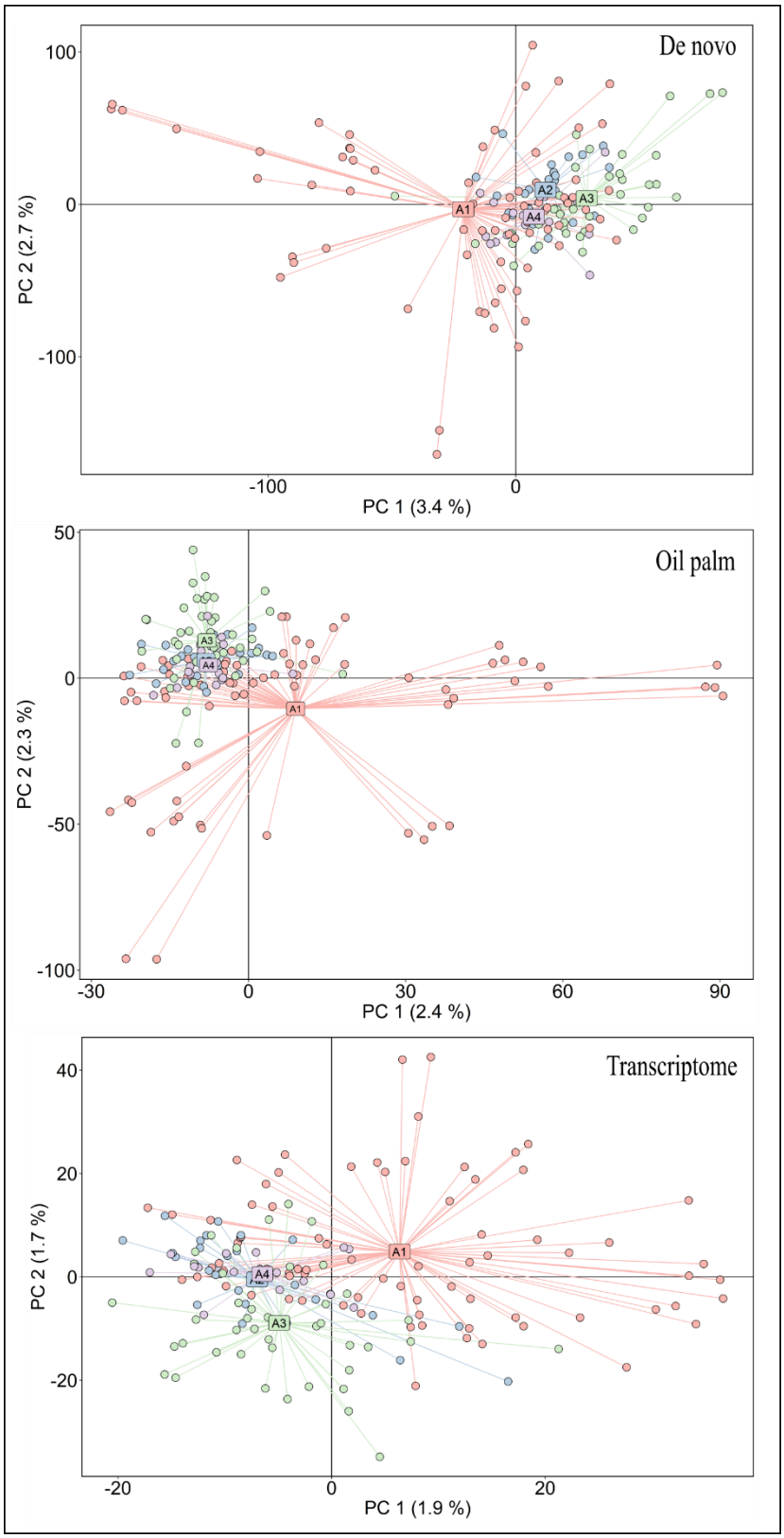

Figure 3. Principal component in the genotypic datasets d*e novo*, oil palm and transcriptome. PCA shows the first two principal components in parentheses in the axis titles. Colors dots represent the four *Acrocomia aculeata* subgroups selected in the four farm areas in this study.

3.3 *Single-trait* GWAS

The three genotypic datasets used in this study showed significant SNPs in the single-trait GWAS for the GLM, MLMM, FarmCPU and BLINK models. The MLM model did not detect significant SNP in any trait or genotypic datasets. The oil palm dataset showed a total of 19 significant SNPs in the NW, OC, FW, PFM and PDM traits; in the models GLM, MLMM, FarmCPU and BLINK (**Table 3**). From them, 10 SNPs occurred in genes regions (**Table 4**). The d*e novo* dataset showed a total of 92 significant SNPs in the DBH, NL, NW, LL, OC, FW, HFM, PFM, EFM, AFM, FDW, HDM, PDM and EDM traits, in the models GLM, MLMM, FarmCPU and BLINK, being that two occured in genes regions (**Table 3 and 5**). To the *de novo* dataset, many of the significative SNP were associated more than once in the same statistical model, for different traits (**Suplemmentary Table 1**). Transcriptome dataset showed one significant SNP in the FDW trait, in the MLMM model, and this SNP had a gene annotation in the *Elaeis guineensis* (**Table 3 and 6**).

For the candidate genees, the marker SNC_025993.1, in the oil palm dataset, was detected in three single-trait models (GLM, FarmCPU and BLINK), for the same loci (LOC105060459), which has a membrane function (**Table 4**). In addition, the 620475_46 SNP marker, in the *de novo* dataset was associated with 20 candidate genes in different species, all of them having 100% of similarity (data not shown). **Table 5** shows three of these 20 candidate genes, being similar to the *Elaeis guineenses* and *Phoenix dactylifera*, which are phylogenetically closer to macaúba. In all the genotypic dataset used, candidate genes showed similarity with the species *Elaeis guineensis* (100%), *Phoenix dactylifera* (100%) and *Cocos nucifera* (91.3%); and presented function as protein transport, DNA repair, metal ion transport, maturation of ribosome subunit and transmembrane transportation.

3.5 *Multi-trait* GWAS

In the multi-trait GWAS the MSTEP model detected a total of 19 markers associated with traits in all the genotypic dataset studied (**Table 3**). From them, similarities analysis in the Blast2Go software detected three candidate genes in the oil palm dataset to eight trait combination, and one in the transcriptome dataset to two trait combination (**Table 3 and 7**). To the oil palm dataset, SNC_025993.1, SNC_025995.1, and SNC_026000.1 markers had similarities with *Elaeis guineensis* genes in eight trait combinations. The SNC_025993.1 marker, position 60355952, showed association with

the traits combinations H versus NW, and DBH versus NW, having membrane function. Furthermore, this marker was also detected in three single-trait GWAS models (**Table 4**). The second marker, SNC_025995.1, position 17709887, had association with the traits LL versus OC, and LN versus OC, having also been detected in the single-trait GWAS for the OC trait (**Table 4**). In addition, the SNC_026000.1 marker, position 5081018, had association with the traits LN versus HFM, NN versus HFM, NN versus FW, and NN versus HDM. For the transcriptome dataset the STRINITY9279 marker, associated with the traits FDW versus OC, and FDW versus HDM, showed similarity with the *Elaeis guineenses* uncharacterized protein (**Table 7**). To a better visualization of the significant SNPs detected in the multi-trait model, we added in the supplementary information the manhattan plots obtained in all the single-trait GWAS. In these graphs, it is possible to observe red dots, which represent the significant SNPs detected in the multi-trait GWAS. Some of these red dots are located in regions that are not significant for the considered *p-value*. However, other red dots are localizated in the same significant SNPs dots associated with the single-trait GWAS, confirming the real association of the SNP with the studied traits (**Supplementary material**).

Table 3. Single-trait and multi-trait GWAS in the genotypic dataset used in this work: *de novo*, oil palm and transcriptome. Total of significative SNP shows the number of significative SNP associated with the studied traits, considering the duplicated SNPs. Total of candidate genes show the number of SNP similar with genes in different species.

GWAS	Genotypic dataset	Total of significative SNP	Total of candidate genes
Single-trait	De novo	92	2
Single-trait	Oil palm	19	10
Single-trait	Transcriptome	1	1
Multi-trait	De novo	6	0
Multi-trait	Oil palm	11	3
Multi-trait	Transcriptome	2	1

Figure 4. Significant SNPs detected by different single-trait (GLM, MLMM, FarmCPU, BLINK) and multi-trait (MSTEP) GWAS models in each genotypic dataset. Venn Diagramm shows the unique and shared significative SNP in the de novo (A), oil palm (B) and transcriptome (C) datasets. The figure represent the total number of SNP markers in each dataset without the duplicated SNPs in the de novo dataset.

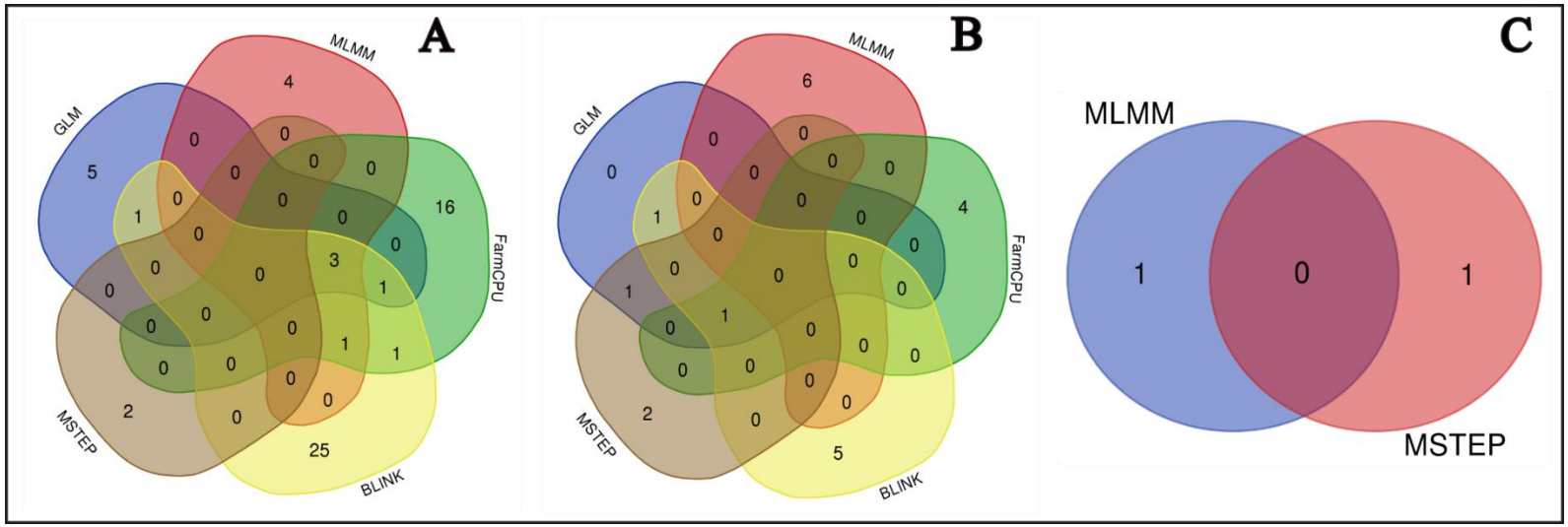

Table 4. Candidate genes from the oil palm dataset significantly associated with NW, OC, FW and PFM traits by diferent single-trait GWAS models.

Trait	Single-trait GWAS	Candidate gene	Loci	Chr.	Position (bp)	*P* value	MAF	Gene Annotation	Scientific Taxonomy	Function
NW	GLM FarmCPU BLINK	SNC_025993.1	LOC105060459	1	60355952	0.0000029 0.0000000 0.0000000	0.44	Uncharacterized protein	*Elaeis guineensis*	Membrane
	BLINK	SNC_025997.1	LOC105045291	5	18357165	0.0000034	0.07	Ribosome biogenesis protein NSA2 homolog	*Elaeis guineensis*	Maturation of 5.8S rRNA; Maturation of LSU-rRNA; Nucleolus; Preribosome
OC	GLM	SNC_025995.1	LOC105041056	3	17709887	0.0000023	0.026	Zinc finger protein VAR3, chloroplastic	*Elaeis guineensis*	Metal ion binding
	FarmCPU	SNC_025999.1	LOC105048521	7	11453414	0.0000003	0.036	Uncharacterized protein	*Elaeis guineensis*	Protein transport
	FarmCPU	SNC_026004.1	LOC105054947	12	15076493	0.0000039	0.31	Mitochondrial import inner membrane translocase	*Elaeis guineensis*	Protein transport; Mitochondrial inner membrane;

								subunit Tim13 isoform X2		
FW	BLINK	SNW_011550986.1	LOC105032941	-	1057	0.0000018	0.18	Glutathione reductase, cytosolic isoform X2	*Elaeis guineensis*	Glutathione metabolic process; Cell redox homeostasis; Cellular oxidant detoxification; Flavin adenine dinucleotide
	BLINK	SNW_011552849.1	-	-	2295	0.0000001	0.14	GDSL esterase/lipase LIP-4	*Elaeis guineensis*	Hydrolase activity
PFM	MLMM	SNC_025993.1	LOC105034069	1	67158358	0.0000007	0.08	Extracellular ribonuclease LE-like	*Elaeis guineensis*	RNA phosphodiester bond hydrolysis, endonucleolytic; Ribonuclease T2 activity
	MLMM	SNC_025997.1	LOC105044859	5	8305530	0.0000019	0.18	Putative zinc transporter At3g08650	*Elaeis guineensis*	Metal ion transport; Transmembrane transport; Plasma membrane

	MLMM	SNC_026005.1	LOC105056693	13	24632742	0.0000003	0.02	20 kDa chaperonin, chloroplastic	*Elaeis guineensis*	Protein folding; Positive regulation of superoxide dismutase activity; ATP binding;

NW: needles width, OC: oil content, FW: fruit weight, PFM: pulp fresh mass.
[a] Some of functions from the candidate genes

Table 5. Candidate genes from the *de novo* dataset significantly associated with EFM and HDM traits by diferent single-trait GWAS models

Trait	Single-trait GWAS	Candidate gene	Loci	Chr.	Position (bp)	*P* value	MAF	Gene Annotation	Scientific Taxonomy	Function
EFM	BLINK	568099_49	EHA8587225.1	UN	39766911	0.0000001	0.38	Hypothetical protein gb|EHA8587225.1|	*Cocos nucifera*	
HDM	GLM	620475_46	LOC105052613	UN	43433228	0.0000021	0.05	Protein reduced wall acetylation 4 isoform X1 and X2	*Elaeis guineensis*	Carbohydrate metabolic process; Transferase activity; Golgi apparatus;
			LOC105039974	UN	43433229	0.0000021	0.05	Protein reduced wall acetylation 4	*Elaeis guineensis*	

		LOC103719307	UN	43433230	0.0000021	0.05	Protein reduced wall acetylation 4-like isoform X1, X2, X3, X4 e X5	*Phoenix dactylifera*		

EFM: endocarp fresh mass; HDM: husk dry mass.

Table 6. Candidate gene from the transcriptome dataset significantly associated with FDW trait by MLMM *single-trait* GWAS model

Trait	Single-trait GWAS	Candidate gene	Loci	Chr.	Position (bp)	*P* value	MAF	Gene Annotation	Scientific Taxonomy	Function
FDW	MLMM	STRINITY31179	LOC105037057	1	2650	0.0000045	0.49	Uncharacterized protein	*Elaeis guineensis*	DNA repair; DNA binding; Catalytic activity;

FDW: fruit dry weight

Table 7. Multi-trait GWAS model MSTEP

Dataset	Trait	Candidate gene	Loci	Chr.	Position (bp)	ProbF	Gene Annotation	Scientific Taxonomy	Function
Oil Palm	H and NW DBH and NW	SNC_025993.1	LOC105060459	1	60355952	0.00000086839 0.0000084689	Uncharacterized protein	*Elaeis guineensis*	Membrane
	LL and OC LN and OC	SNC_025995.1	LOC105041056	3	17709887	0.00000031504 0.0000012682	Zinc finger protein VAR3, chloroplastic	*Elaeis guineensis*	Metal ion binding
	LN and HFM	SNC_026000.1	LOC105049570	8	5081018	0.0000053214			-

	NN and HFM					0.00000001784	UDP-N-acetylglucosamine peptide N-acetylglucosaminyltransferase 110 kDa subunit isoform X1	*Elaeis guineensis*	
	NN and FW					0.00000017972			
	NN and HDM					0.0000019697			
Transcriptome	FDW and OC	STRINITY9279	LOC105039939	1	4262	0.0000011	Uncharacterized protein	*Elaeis guineensis*	-
Transcriptome	FDW and HDM					0.0000412			

H: height, DBH: diameter of the stem at breast height, LN: leaves number, LL: leaves length, NN: needles number, NW: needles width, OC: oil content, FW: fruit weight, HFM: husk fresh mass, FDW: fruit dry weight, HDM: husk dry mass.

4. DISCUSSION

Acrocomia is a genus from a neotropical palms distributed in almost all tropical and subtropical America, there are nine recognized species in the genus: *A. aculeata*, *A. totai*, *A. hassleri*, *A. glaucescens*. *A. emensis*; *A. intumescens*; *A. media*, *A. crispa* and *A. corumbaensis* (Lima et al., 2018; Lorenzi et al., 2010; Diaz et al., 2021). From them, macaúba (*Acrocomia aculeata*) has stood out in recent years due to its high percentage of oil content in the pulp and almonds of its fruit, awakening economic interests in the industry due to its capability to supply oil with the same productivity as the oil palm (*Elaeis guineenses*). To our knowledge, this is the first work using genome-wide association study in *Acrocomia aculeata*. Understanding macaúba genetic architecture is an initial step towards its domestication and commercial planting, since markers associated with traits of economic interest can help the selection of promising genotypes in breeding programs.

4.1 Importance of analyzed traits to *Acrocomia aculeata* domestication

The data used in this work were collected from a natural population located inside a private rural propriety, being that the palms are growing in a integrated Crop-Livestock-Forest systems. Obtaining and analyzing macaúba phenotypic data allows to understand the interest traits for the pre-breeding and direct the way for the future commercialization of this palm. In this study, the likelihood ratio test, performed for each of the phenotypic traits, detected significant differences among genotypes for all of them ($p < 0.001$), revealing levels of genetic variability within the *Acrocomia aculeata* population evaluated (**Table 1**). Furthermore, the ranges of the mean values within each trait studied showed a large interval. Once macaúba is a non-domesticated palm, this occurred as a consequence of sampling a *Acrocomia aculeata* natural population. The study did not have an experimental design, which means that the age of the macaúba genotypes was unknow. For example, the height (H) and diameter of the stem at breast height (DHB) are traits that could not be obtained in plants with the same age. Therefore, it is important to bear in mind that a genotype with low height, compared with another, may be low due to its age, and not due to its trait. This is one of the great challenges in studying perennial and native populations. In macaúba, the first flowering occurs in the fifth or sixth year of the palm tree (Teixeira, 2019). Furthermore, fruit development is slow and can last up to 62 weeks after anthesis, equivalent to 14.3 months (Mazzottini-dos-Santos et al., 2015; Montoya et al., 2016). In this way, evaluating native populations becomes an alternative

to studying traits of interest to the biofuel market, aiming to select superior genotypes to start the pre-breeding. Thus, instead of the breeder collecting random fruits from native populations (with good and poor genotypes), he can start the breeding cycle with superior genotypes, saving time. Even without an experimental design, the value of the CV ranged from 22.38% to 5.75%, highlighting good accuracy in obtaining phenotypic data. The percentage of OC ranged of 17.47% to 60.80% in the macauba population; and according to Costa et al. (2018), genetic gains for oil yield per plant are estimated at 61% when superior genotypes are selected. Additionally, H, LL, OC, FW, PFM, EFM, HDM, EDM and ADM traits showed heritability values higher than 80%, evidencing that the performance of breeding cycles in this selected population tends to produce genetic gains in these traits. Other studies have been reported high mean heritability values for various morphological traits in *Acrocomia aculeata* (Berton et al., 2013; Domiciano et al., 2015; Coser et al., 2016; Rosado et al., 2019).

Phenotypic correlation is the result of genetic and environmental causes, while the genotypic correlation corresponds to the genetic portion of the phenotypic correlation. In this case, the genotypic correlation is used in data analysis to guide breeding programs because it is inheritable (Cruz, Regazzi e Carneiro, 2012). Correlation btween traits are of a big interest in plant breeding once they enable a indirect selection of multi-traits (Costa et al., 2018; Domiciano et al., 2015). Furthermore, high values of genetic correlation may indicate the presence of linked genes or in pleiotropy (Fernandes, 2021). In this study, we found high phenotypic and genetic correlations values between the fruit traits as FW with HFM, PFM and PDM; and FDW with PFM, HDM and PDM (**Figure 3**). In the same way, Costa et al. (2018) found highest positive genotypic correlations between macauba fruit traits, as dry matter from endocarp and epicarp, endocarp and kernel, and pulp and oil yield per plant. Ciconini et al., (2013) and Conceição et al. (2015) found positive correlation in pulp oil content with total fruit oil, but a negative correlation of pulp oil content with the mesocarp's proportion of the whole fruit. For OC and NN, null and negative values were observed for all the pairwise traits tested in both correlations studied. Negative and low correlation values were found in other macauba studies: i) between stem diameter at breast height with the height of the first spathe (Coser et al., 2016); ii) between oil content and oil yield per plant (Costa et al., 2018); and iii) between biometric fruit traits and oil content (Ciconini et al., 2013). Further information is needed on selection criteria of the corresponding correlation between traits to assist the breeding of commercial varieties of macauba (Vargas-Carpintero et al. 2021).

Statistical analysis of phenotypic data allows to be carried out a first selection of superior plants, aiming the begining of breeding cycles to obtain a more productive population. Moreover, the analysis of genetic diversity of macaúba is crucial to guide the selection of the most promising materials for crop use, to maximize genetic gains, and to more effectively contribute to the creation of commercial cultivars (Diaz et al., 2021). The three genotypic datasets used in this study showed high values of observed heterozygosity and low values of inbreeding, indicating genetic variability in the population, which enables the beginning of breeding cycles of this species. A ideotype model of the improved macaúba was proposed by Berton (2013). In it, the desirable traits to achieve are: early flowering, low heigth, high production of fruits and oil content, fewer or absence of thorns, indehiscent fruits and overlapping bunches. Therefore, the study of macaúba traits is important to assist in the plant breeding in the selection of potential plants that encompass the traits to be added to the ideotype demanded by the market.

Absence of a reference genome in *Acrocomia aculeata*

Acrocomia aculeata does not have a reference genome. Thus, different strategies were proposed in this work to performe SNP calling: using the *de novo* pipeline (Catchen et al., 2013); the reference genome of *Elaeis guineensis* var *tenera*, and using the transcriptome of *Acrocomia aculeata* (Bazzo et al., 2018). *Elaeis guineensis* reference genome was used in this work due to the phylogenetic proximity to *Acrocomia aculeata* (Bazzo et al., 2018). 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 tree, *Acrocomia aculeata* had a proximity of *Elaeis guineensis* (Francisconi et al., 2023). The three genotypic datasets used in this study also showed different results in the genetic diversity and PCA analyses, as well as in the GWAS. The number of individuals and SNP markers used in the genetic diversity analyses were different in each genotypic dataset because the SNP calling and the quality control step (**Table 2**).

Population genetic parameters in macauba individuals from the oil palm and transcriptome dataset presented higher heterozygosity values than the population from *de novo* dataset (**Table 2**). Therefore, the inbreeding coefficient was higher in the population of *de novo* dataset than in the other two. These values indicate that the heterozygosity and inbreeding coefficient values of the studied macauba populations depended on the strategy used in the SNP calling. Diaz et al. (2021) studied genetic diversity in the Acrocomia genus using genome wide SNP and the *de novo* pipeline. They also reported

low heterozygosity values (0.031). In the *de novo* pipeline from the software Stacks (Catchen et al., 2011), the identification of SNP used the *ustacks* module to identify groups of putatively homologous reads for each studied macauba genotype, separately. An SNP model in *ustacks* checks each locus at each nucleotide position for polymorphisms. Subsequently, a locus catalog was built using the cstacks module, allowing a maximum of two differences between stacks from different individuals. Because the Stacks methodology cited above, the population genetic parameters in *Acrocomia aculeata* from *de novo* dateset showed low heterozygosity values and higher inbreeding coeficiente values when compared to the other datasets. On the contrary, compared to the oil palm and transcriptome dataset, higher heterozygosity values and lower inbreeding coefficient values were observed. This occured because the *Elaeis guineensis* reference genome used in our study present higher genetic variability in its alleles, than the genome of the *de novo* pipeline. Even though Acrocomia and Elaeis are close genera, when you compare two different species divergent alleles are expected from their evolutionary history. The transcriptome dataset was generated by the alignment step with the transcriptome sequencing of *Acrocomia aculeata* (Bazzo et al., 2018). Bazzo and collaborators reported the identification and validation of 145 macauba EST-SSR markers from different tissues and individuals using next-generation sequencing. The study was performed using tissues of leaves, leaf sheaves, roots, bulbs, fruit (mesocarp and endosperm) and male and female flowers from the macaúba for RNA isolation and transcriptome sequencing. The reads originating from mRNA libraries were mapped against the African oil palm genome (Singh et al., 2013). Results indicated a 80.7% in transferability rates of macauba in African palm. It is important to point out that the EST-SSR markes are multiallelic, and compared to biallelic SNPs they could bias genetic variability.

In addition, PCA revealed that individuals from subgroup 1 are slightly separated from subgroup 2, 3, and 4 (**Figure 3**). However, among the genotypic datasets it is possible to see that subgroups 2, 3 and 4 presented variations in their groupings, showing that SNP calling strategies may lead to different interpretations of the population subgroups. In the transcriptome dataset, the subgroup 3 is a little distant from subgroups 2 and 4, while in the oil palm dataset the subgroups 2, 3 and 4 are tightly grouped.

Results showed here indicate that the SNP calling strategies is important to provides a direction in understanding how the macauba population studied here is behaving in terms of diversity and genetic architecture.

4.2 Single-trait and multi-trait GWAS in macauba

Using single-trait and multi-trait GWAS models we identifyed SNP markers located in gene regions related to different phenotypic traits involved in fruit production and fruit pulp oil content of a macaúba population. The results related to the genotype-to-phenotype association in these traits could be relevant for understanding the genetic architecture of this neotropical palm. Moreover, the genetic mapping and molecular characterization of these genes that contribute to the variation of complex traits has the potential to facilitate genome-assisted breeding for crop improvement (Stich & Melchinger, 2010), and in the case of macaúba can possibility the beginning of its domestication. The elucidation of its genetic information could still speed up the implementation of pre-breeding programs to select superior genotypes for traits involved with the oil production of *Acrocomia aculeata*.

The total number of significant SNPs detected in the single-trait and multi-trait GWAS was greater than the number of candidate genes (**Table 3**). This result shows that false positives could have been detected by the statistical models used, as well as highlighting the importance of obtaining the reference genome for the macaúba. The single-trait GWAS revealed significant genomic regions in the three dataset used in this study, associated with the following traits: DBH, LL, NL, NW, OC, FW, HFM, PFM, EFM, AFM, FDW, HDM, PDM and EDM. For the traits H and LN, no genomic regions were found in the search for similarity in the software Blast2GO (**Table 4, 5 and 6**). This result may have occurred because the phenotypic data came from native palms, whose lifetime is unknown. Therefore, it was not possible to control the height of the plants taking their age into account, as already discussed in the topic above. For the LN trait, phenotypic data were obtained through manual counting. There was no removal of palm leaves to account for their total quantity, making it difficult to control the collection of this data.

Statistical models have been developed for GWAS to distinguish real associations from false positives caused by population structure and linkage disequilibrium. Some of the first GWAS models were single-locus, single-trait GWAS models created to implement the covariates along with kinship matrices (Yu et al., 2006). Later, multi-trait GWAS models were developed to analyzed multiple traits simultaneously, mainly to increase statistical power and identify pleiotropic loci (Porter and O'Reilly, 2017, Fernandes et al., 2021). In our study, 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 detected for different trait combination (**Supplementary Table 1**). Similar results were observed in the multi-trait models (**Table 7**). Moreover, for the multi-trait model, four candidate genes (SNC_025993.1, SNC_025995.1, SNC_026000.1, STRINITY9279) stood out in the phenotype-genotype association between different combinations of traits. Among these traits, NW, OC, NN, HFM, HDM, FDW appeared more than once. Furthermore, the SNP markers SNC_025993.1 and SNC_025995.1, detected in the multi-trait model, were also detected in the single-trait models. The supplementary material presents the significative SNPs detected by the single-trait and multi-trait (red dots) GWAS. On page 36, for example, for the NW trait, the same significative SNP was detected by the BLINK and MSTEP models used in our study. These results demonstrate the certainty of the presence of these loci in macaúba genome, reducing the chance of these SNP markers being false positives. In this context, Fernandes et al. (2021) suggest that single-trait and multi-trait GWAS should both be used to infer whether or not causal mutations underlying peak GWAS associations are pleiotropic. However, the authors discuss about the importance to note that multivariate GWAS could not distinguish between QTNs in LD and a single pleiotropic QTN, underlying validation analyzes of the significant SNPs detected by the GWAS models are necessary.

4.3 Future perspectives

GWAS has been performed on many domesticated crop populations even in the oil palm, for different and important economic traits (Ithinin et al., 2017; Babu et al., 2020 and 2021). GWAS has been used even to investigate pleiotropy (Wisser et al., 2011; Schaid et al., 2016; Rice et al., 2020). In this situation, phenotype-genotype trait association is analyzed using a statistical design, because the genotypes in the studied population are already known and sometimes, well established. Furthermore, in general, domesticated species have an reference genome, which possibility to understand their genetic architecture from GWAS studies. In the case of macaúba, the species does not have already a reference genome, and the similarity of significant SNPs has to be looked for in genomes of related species. Additionally, the positions of the SNPS in the genome will occur accordingly with the genome of the species used as reference. In our study, genotypes were selected in their natural habitat with the aim of starting the understand the genetic architecture of macauba, to opening ways for new studies to be carried out. This is the first work using GWAS in *Acrocomia aculeata*, and the significant SNPs

detected here can be validated in a future study, in order to allow marker-assisted selection to support the selection of traits of interest in native plants in other populations.

CONCLUSIONS

We provide new insights on genomic regions that mapped candidate genes involved in macaúba oil production phenotypes. These potential candidate genes need to be confirmed for future targeted functional analyses and multi-trait associations need to be scrutinized to investigate the presence of pleiotropic or linked genes. Associated markers to the traits of interest may be valuable resources for the development of marker-assisted selection in macaúba for its domestication and pre-breeding.

ACKNOWLEDGEMENTS

We would like to thank graduate students Luiz Henrique Berton, Gabriel Alves, Elivelton Alves, and technician Evandro Coelho, from the Campinas Agronomic Institute and graduate students from the Luiz de Queiroz College of Agriculture, Marcela Barbosa, Laecio Sampaio and Marcelo Almeida, for their assistance in obtaining phenotypic data and fruit biometry. We would also like to thank Bárbara Regina Bazzo, Lucas Miguel de Carvalho and Marcelo Falsarella Carazzolle who provided the macaúba transcripts for the SNP calling.

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).

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The first draft of the manuscript was written by Evellyn Giselly de Oliveira Couto, and all authors read and approved the final manuscript prior to publication. The contribution of each author is presented below:

Evellyn Giselly de Oliveira Couto: Data collect, Genotypic data obtation, Methodology, Data mining analyses, Data curation, Conceptualization, Investigation, Writing, Review & Editing.

Jonathan Morales-Marroquín: Data collect, Genotypic data obtation, Methodology, Investigation, Review & Editing.

Alessandro Alves-Pereira: Bioinformatics analyses, Data mining analyses, Validation.

Samuel B. Fernandes: Conceptualization, Methodology, Data mining analyses, Validation, Review & Editing.

Carlos Augusto Colombo: Data collect, Resources, Investigation, Funding acquisition.

Joaquim Adelino de Azevedo Filho: Data collect, Resources, Investigation.

Maria Imaculada Zucchi: Conceptualization, Resources, Investigation, Review & Editing, Funding acquisition, Supervision.

REFERENCES

Allen J.M., Boyd B., Nguyen N., Vachaspati P., Warnow T., Huang D.I., Grady P.G.S., Bell K.C., Cronk Q.C.B., Mugisha, L., Pittendrigh B.R., Soledad Leonardi M., Reed D.L., & Johnson, K.P. Phylogenomics from whole genome sequences using aTRAM (2017). Systematic Biology, 66, 786–798.

Abreu A.G., Priolli R.H.G, Azevedo-Filho J.A., Nucci S.M., Zucchi M.I., Coelho R.M., Colombo A.C. The Genetic Structure and Mating System of Acrocomia Aculeata (Arecaceae) (2012). Genetics and Molecular Biology, 35(1), 116–121.

Babu B.K., Mathur R.K., Ravichandran G., Anita P., Venu M.V.B. 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 (2020). Genomics, 112(1),1011-1020.

Babu B.K., Mathur R.K., M.V.B.V., Shil S., et al. Genome-wide association study (GWAS) of major QTLs for bunch and oil yield related traits in Elaeis guineensis L (2021). Plant Science, 305, 110810.

Bazzo B.R., Carvalho L.M., Carazzolle M.F., Pereira G.A.G., Colombo C.A. Development of Novel EST-SSR Markers in the Macaúba Palm (Acrocomia Aculeata) Using Transcriptome Sequencing and Cross-Species Transferability in Arecaceae Species (2018). BMC Plant Biology, 8(1).

Berton, Luiz Henrique Chorfi. Avaliação de populações naturais, estimativas de parâmetros genéticos e seleção de genótipos elite de macaúba (*Acrocomia aculeata*) (2013) Tese. Campinas.

Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y.,

& Buckler E.S. TASSEL: Software for association mapping of complex traits in diverse samples (2007). Bioinformatics, 23, 2633–2635.

Browning, B.L., Tian X., Zhou Y., Browning S.R. Fast two-stage phasingof large-scale sequence data (2021). American Journal of Human Genetics108(10), 1880-1890.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. Stacks: Building and genotyping loci de novo from shortread sequences (2011). G3: Genes, Genomes, Genetics, 1(3).

Chen H., Boutros P.C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R (2011). BMC Bioinformatics, 12:35.

Cichonska A., Rousu J., Marttinen P., Kangas A. J., Soininen P., Lehtimäki T., et al. metaCCA: summary statistics-based multivariate metaanalysis of genome-wide association studies using canonical correlation analysis (2016). Bioinformatics 32, 1981–1989.

Ciconini G., Favaro S.P., Roscoe R., Miranda C.H.B., Tapeti C.F., Miyahira M.A.M., Bearari L., Galvani F., Borsato A.V., Colnago L.A., Naka M.H. Biometry and Oil Contents of Acrocomia Aculeata Fruits from Cerrados and Pantanal Biomes in Mato Grosso Do Sul, Brazil (2013). Industrial Crops and Products, 45, 208–214,

Coimbra M.C., 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) (2012). Journal of the Science of Food and Agriculture, 92(3),679–684.

Colombo C.A., Berton L.H.C., Diaz B.G., Ferrari R.A. Macauba: A Promising Tropical Palm for the Production of Vegetable Oil (2018). OCL, 25(1), 1-9.

Conesa A., Gotz S., Garcia-Gomez J.M., Terol J., Talon M., Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research (2005). Bioinformatics, 21: 3674–3676.

Coser S.M., Motoike S.Y., Corrêa T.R., Pires T.P., Resende M.D.V. Breeding of Acrocomia aculeata using genetic diversity parameters and correlations to select accessions based on vegetative, phenological, and reproductive characteristics (2016). Genetic Molecular Research, 15(4).

Covarrubias-Pazaran G. (2016). “Genome assisted prediction of quantitative traits using the R package sommer.” PLoS ONE, 11, 1-15.

Cruz C.D., Regazzi A.J, Carneiro P.C.S. Modelos biométricos aplicados ao melhoramento genético. 4. ed. Viçosa, MG: Ed. UFV, 2012.
da Conceição L.D.H.C.S., Antoniassi R., Junqueira N.T.V. et al. Genetic diversity of macauba from natural populations of Brazil (2015). BMC Res Notes 8, 406.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., et al. The variant call format and VCFtools (2011). Bioinformatics, 27(15), 2156–2158.

Doyle J.J., Doyle J.L. Isolation ofplant DNA from fresh tissue (1990). Focus, 12(13), 39–40.

Diaz B.G., Zucchi M.I., Alves-Pereira A., Almeida C.P.A., Moraes A.C.L., Vianna S.A., Azevedo-Filho J., Colombo C.A. Genome-wide SNP analysis to assess the genetic population structure and diversity of Acrocomia species (2021). PLoS ONE, 16(7): 1-24.

Domiciano G.P., Alves A.A., Laviola B.G., da Conceição L.D.H.C.S. Genetic parameters and diversity in progenies from macaw palm based on morphological and physiological traits [Parâmetros genéticos e diversidade em progênies de macaúba com base em características morfológicas e fsiológicas] (2015). Ciência Rural 45(9), 1599–1605.

Falconer D.S., Mackay T.F.C. Introduction to quantitative genetics. 4 ed. ed. New York, NY: Longman Group Limited: Edinburgh, 1996.

Francisconi A.F., Marroquín J.A.M., Cauz-Santos L.A. *et al.* Complete chloroplast genomes of six neotropical palm species, structural comparison, and evolutionary dynamic patterns (2023). Scientific Reports, 13, 20635.

François O., Caye K. Naturalgwas: An R package for evaluating genomewide association methods with empirical data (2018). Molecular Ecology Resources, Special issue: Association mapping in natural populations.

Fernandes S.B., Zhang K.S., Jamann T.M. and Lipka A.E. How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy? (2021). Frontiers in Genetics, 11:602526.

Fernandes S.B., Casstevens T.M., Bradbury P.J., Lipka A.E. A multi-trait multi-locus stepwise approach for conducting GWAS on correlated traits (2022). Plant Genome, 15:e20200

Filzmoser P., Ruiz-Gazen A., and Thomas-Agnan C.: Identification of local multivariate outliers. Submitted for publication, 2012.

Furlotte N.A., & Eskin E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model (2015). Genetics, 200, 59–68.

Garrison E., Marth G. Haplotype-based variant detection from short-read sequencing (2012). arXiv preprint arXiv:1207.3907.

Goudet J. Hierfstat, a package for R to compute and test hierarchical F-statistics (2005). Molecular Ecology Notes, 5: 184-186.

Heberle, H., Meirelles, G.V., da Silva, F.R. *et al.* InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams (2015). BMC Bioinformatics 16, 169.

Hiane P.A., Filho M.M.R., Ramos M.I.L, Macedo M.L. Bocaiuva, Acrocomia aculeata (Jacq.) Lodd., pulp and kernel oils: characterization and fatty acid composition (2005). Brazilian journal of food technology, p. 256–259.

Huang M., Liu X., Zhou Y., Summers R.M., Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions (2019). Gigascience, 8:1–12.

Ithnin M., Xu Y., Marjuni M. et al. Multiple locus genome-wide association studies for important economic traits of oil palm (2017). Tree Genetics & Genomes 13, 103.

Jombart T. “adegenet: a R package for the multivariate analysis of genetic markers.” (2008). Bioinformatics, 24, 1403-1405.

Jombart T., Ahmed I. “adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.” (2011). Bioinformatics, 27(21), 3070-1.

Joo J.W.J., Kang E.Y., Org E., Furlotte N., Parks B., Hormozdiari F., et al. Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure (2016). Genetics 204, 1379–1390.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., Subgroup 1000 Genome Project Data Processing. The Sequence Alignment/Map format and SAMtools (2009). Bioinformatics 25, 2078–2079.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 2013. arXiv:1303.3997v2.

Lima N.E., Carvalho A.A., Meerow A.W. *et al.* A review of the palm genus *Acrocomia*: Neotropical green gold (2018). Organisms Diversity & Evolution, 18, 151–161.

Liu X., Huang M., Fan B., Buckler E.S., Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome- wide association studies (2016). PLoS Genet,12:e1005767

Lorenzi H., Noblick L., Kahn F., & Ferreira, E. Flora Brasileira: Arecaceae (palmeiras). Nova Odessa, SP: Instituto plantarum, 2010.

Mackay T.F.C. Quantitative Trait Loci in Drosophila (2001). Nature Reviews Genetics, 2(1), 11–20.

Malosetti M., Ribaut J.M., Vargas M. et al. A Multi-Trait Multi-Environment QTL Mixed Model with an Application to Drought and Nitrogen Stress Trials in Maize (Zea Mays L.) (2008). Euphytica, 161, 241–257.

Mazzottini-dos-Santos H.C., Ribeiro L.M., Mercadante-Simões M.O. et al. Floral structure in Acrocomia aculeata (Arecaceae): evolutionary and ecological aspects (2015). Plant Syst Evol 301, 1425–1440.

Merrick L.F., Burke A.B., Zhang Z. and Carter A.H. Comparison of Single-Trait and Multi-Trait Genome-Wide Association Models and Inclusion of Correlated Traits in the Dissection of the Genetic Architecture of a Complex Trait in a Breeding Program (2022). Frontiers in Plant Science, 12, 772907

Montoya, S.G., Motoike, S.Y., Kuki, K.N. et al. Fruit development, growth, and stored reserves in macauba palm (Acrocomia aculeata), an alternative bioenergy crop (2016). Planta 244, 927–938.

Oraguzie N.C. et al. Association Mapping in Plants. New York, NY: Springer New York, 2007.

Peterson R.A., Cavanaugh J.E. "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era." (2020). Journal of Applied Statistics, 47(13-15), 2312-2327.

Poland J.A., Rife T.W. Genotyping-by-Sequencing for Plant Breeding and Genetics (2012). The Plant Genome, 5 (3).

Poland J.A., Brown P.J., Sorrells M.E., Jannink J.L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach (2012). PloS one, 7(2), e32253.

Porter H.F., and O'Reilly P. F. Multivariate simulation framework reveals performance of multi-trait GWAS methods (2017). Scientific Reports, 7:38837.

Pritikin J.N., Neale M.C., Prom-Wormley, E.C., Clark, S.L., & Verhulst, B. GW-SEM 2.0: Efficient, flexible, and accessible multivariate GWAS (2021). Behavior Genetics, 51, 343–357.

Quinlan A.R. and Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features (2012). Bioinformatics. 26(6), 841–842.

R Core Team. R: A language and environment for statistical computing (2021). R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rice, B. R., Fernandes, S. B., & Lipka, A. E. Multi-trait genomewide association studies reveal loci associated with maize inflorescence and leaf architecture (2020). *Plant & Cell Physiology*, *61*(8), 1427–1437.

Rosado RDS, Rosado TB, Cruz CD, Ferraz AG, da Conceição LDHCS, Laviola BG Genetic parameters and simultaneous selection for adaptability and stability of macaw palm (2019). Scie HorticAmsterdam 248, 291–296

Scariot A., Lleras E., HAY J.D. Flowering and Fruiting Phenologies of the Palm Acrocomia aculeata: Patterns and Consequences (1995). Biotropica, 27(2), 168.

Schaid D.J., Tong X., Larrabee B., Kennedy R.B., Poland G.A., Sinnwell J.P. Statistical methods for testing genetic pleiotropy (2016). Genetics 204, 483–497.

Silva J.C., Barrichelo L.E.G., Brito J.O. Endocarpos de Macaúba e de Babaçu comparados a madeira de Eucaliptus grandis na produção de carvão vegetal (1986). IPEF, 34, 31–34.

Singh R., Ong-Abdullah M., Low E-TL., Manaf M.A.A., Rosli R., Nookiah R., et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds (2013). Nature, 500:335–9.

Stich B., Melchinger A.E., An introduction to association mapping in plants (2010). CABI Reviews, 5, No. 039

Teixeira L.C. Potencialidades de oleaginosas para produção de biodiesel (2005). Informe Agropecuário, 26, 18–27.

Teixeira N., Melo J.C.S., Batista L.F., Paula-Souza J., Fronza P., Brandão M.G.L. Edible fruits from Brazilian biodiversity: a review on their sensorial characteristics versus bioactivity as tool to select research (2019). Food Res Int 119, 325–348.

Vargas-Carpintero, R., Hilger T., Mössinger J., Souza R.F., Armas J.C.B.A., Tiede K., Lewandowski I. *Acrocomia* spp.; negleted crop, ballyhooed multipurpose palm or fit for the bioeconomy? A review (2021). Agronomy for sustainable development, 41 (6), 75.

Wang Q., Tian F., Pan Y., Buckler E. S., and Zhang Z. A SUPER powerful method for genome wide association study (2014). PLoS ONE, 9, e107684.

Wang, J. & Zhang, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction (2021). Genomics Proteomics Bioinformatics 19, 1–12.

Wisser, R. J., Kolkman, J. M., Patzoldt, M. E., Holland, J. B., Yu, J., Krakowsky, M., et al. Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene (2011). PNAS. 108, 7339–7344.

Yu J. et al. A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness (2006). Nature Genetics, 38(2), 203–208.

Zhang Z., Ersoz E., Lai C. Q., Todhunter R. J., Tiwari H. K., Gore M. A., et al. Mixed linear model approach adapted for genome-wide association Studies (2010). Nature Genetics, 42, 355–360.

Zhou X., Stephens M. Efficient Multivariate Linear Mixed Model Algorithms for Genome-Wide Association Studies (2014). Nature Methods, 11(4), 407–409.

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11, 407–409.

Zhu C. et al. Status and Prospects of Association Mapping in Plants (2008). The Plant Genome, 1(1), 5.

SUPPLEMENTARY INFORMATION

Supplementary Table 1. Same SNP markers associated with different traits in the same *single-trait* GWAS model for the *de novo* dataset. Chromosome (Chr.), position of the marker, *p* value, minnor allele frequency (MAF) and effect of each marker.

Dataset	Trait	*Single-trait* GWAS	SNP marker	Chr.	Position (bp)	*P* value	MAF
De novo	FW	GLM	206405	UN	14448334	0.0000022	0.061
De novo	PFM	GLM	206405	UN	14448334	0.0000012	0.061
De novo	FDW	GLM	206405	UN	14448334	0.0000010	0.061
De novo	PDM	GLM	206405	UN	14448334	0.0000002	0.061
De novo	FDW	MLMM	45983	UN	3218785	0.0000000	0.027
De novo	HDM	MLMM	45983	UN	3218785	0.0000000	0.027
De novo	PDM	MLMM	45983	UN	3218785	0.0000003	0.027
De novo	FW	MLMM	179706	UN	12579372	0.0000000	0.044
De novo	PFM	MLMM	179706	UN	12579372	0.0000001	0.044
De novo	FW	MLMM	271071	UN	18974943	0.0000000	0.030
De novo	PFM	MLMM	271071	UN	18974943	0.0000021	0.030
De novo	FDW	MLMM	415478	UN	29083400	0.0000000	0.023
De novo	PDM	MLMM	415478	UN	29083400	0.0000000	0.023
De novo	HDM	FarmCPU	45983	UN	3218785	0.0000000	0.027
De novo	PDM	FarmCPU	45983	UN	3218785	0.0000000	0.027
De novo	PFM	FarmCPU	345096	UN	24156704	0.0000000	0.367
De novo	HDM	FarmCPU	345096	UN	24156704	0.0000010	0.360
De novo	PDM	FarmCPU	345096	UN	24156704	0.0000003	0.367
De novo	PFM	FarmCPU	415478	UN	29083400	0,0006671	0.023
De novo	FDW	FarmCPU	415478	UN	29083400	0,0588715	0.023
De novo	PDM	FarmCPU	415478	UN	29083400	0.0000000	0.023
De novo	EFM	BLINK	39315	UN	2752022	0.0000000	0.030
De novo	EDM	BLINK	39315	UN	2752022	0.0000005	0.030
De novo	LL	BLINK	45983	UN	3218785	0.0000000	0.026
De novo	HFM	BLINK	45983	UN	3218785	0.0000006	0.027
De novo	PFM	BLINK	45983	UN	3218785	0.0000000	0.027
De novo	FDW	BLINK	45983	UN	3218785	0.0000000	0.027
De novo	HDM	BLINK	45983	UN	3218785	0.0000000	0.027
De novo	EFM	BLINK	568099	UN	39766911	0.0000001	0.381
De novo	EDM	BLINK	568099	UN	39766911	0.0000000	0.381
De novo	EFM	BLINK	591967	UN	41437644	0.0000000	0.030
De novo	EDM	BLINK	591967	UN	41437644	0.0000000	0.030