Landscape genomics on Acrocomia aculeata trees aiming breeding selection

Evellyn G. O. Couto
evellyncoutoo@gmail.com
Federal University of Viçosa
Lucas L. S. Peixoto
Federal University of Goiás
Miriam Suzane Vidotti
Federal University of Goiás
Kacilda Naomi Kuki
Federal University of Viçosa
Diego Ismael Rocha
Federal University of Viçosa
Sérgio Yoshimtsu Motoike
Federal University of Viçosa
Maria Imaculada Zucchi
University of São Paulo
Rafael T. Resende
Federal University of Goiás

Research Article

Keywords: Macauba, genomic selection, sustainable oil production, environmental adaptation

Posted Date: May 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9053189/v1

Additional Declarations: No competing interests reported.

Abstract

Landscape genomics identifies genomic regions and alleles associated with environmental adaptation by correlating allele frequencies with environmental variables, providing insights into genotype-environment associations and local adaptation patterns. When integrated with genomic selection, this approach can improve the accuracy of predicting performance in target environments and support the selection of genotypes with high genetic merit and adaptive potential. Here, we present an integrative framework combining landscape genomics and genomic selection to support conservation and breeding strategies in *Acrocomia aculeata* (Jacq.) Lodd. ex Mart. By integrating spatial and genomic information, we characterized patterns of genetic variation and identified superior native trees with potential for genetic improvement. Three alternative SNP-calling strategies were evaluated: a *de novo* pipeline, the *Elaeis guineensis* reference genome, and the *A. aculeata* transcriptome. These datasets were used to assess the predictive ability of genomic selection models for pulp dry mass, kernel dry mass, and oil content. To validate the models, the leave-one-out cross-validation and site-specific validation of the phenotypic traits were performed. Across SNP-calling strategies and validation schemes, genomic prediction achieved moderate to high predictive accuracies for all traits, demonstrating the feasibility of genomic selection in this species. Spatial analyses revealed no strong geographic clustering of superior genetic values, indicating that high-performing genotypes are broadly distributed across the landscape. Overall, our results support seed collection based on individual superior trees and highlight the potential of integrating genomic prediction into macauba breeding and conservation programs.

1. Introduction

Landscape genomics integrates population genetics and landscape ecology to investigate how genetic variation within natural populations is influenced by environmental factors (Manel et al. 2003; Konzen and Zucchi 2020). By focusing on spatial patterns, this approach enables the examination of microevolutionary processes that shape the species under study, providing a clearer understanding of interactions between populations and their environments (Holderegger and Wagner 2006). For *Acrocomia aculeata* (Jacq.) Lodd. ex Mart (popularly known as macauba - in English macaw palm), landscape genomics can reveal how genotypes and environments interact across various agroecological regions. *A. aculeata* is a neotropical oleaginous palm, widely distributed in tropical landscapes, and in Brazil, it forms large populations along a latitudinal gradient (Scariot et al. 1995; Lorenzi et al. 2010; Lima et al. 2018). Its adaptability to diverse environmental conditions makes it an ideal candidate for sustainable cultivation in sensible agricultural systems, such as agroforestry or intercropping systems (Yaqoob et al. 2023). The oil of *A. aculeata*, extracted from both the fruit pulp and seed, holds promising applications in the food, pharmaceutical, and bioenergy industries (Vargas-Carpintero et al. 2021). Combining landscape genomics with genomic prediction studies can assist in selecting superior native trees and can help identify genetic variation patterns that inform more efficient management practices, tailored to local environments, thus maximizing the use of the species' genetic diversity to improve productivity and adaptability to different planting areas (Konzen and Zucchi 2020). Efforts to regulate the cultivation of *A. aculeata*, including identifying the most suitable regions for its planting through agroecological zoning and the Agricultural Climate Risk Zoning (ZARC), are already available in https://mapa-indicadores.agricultura (Resende et al. 2020; Ministério da Agricultura e Pecuária 2026). However, there remains a lack of precise information regarding genotype-environment interactions.

Populations of *A. aculeata* collected from different locations exhibited wide genetic variability, enabling the development of cultivars with higher fruit and oil yields from both pulp and seed (Díaz et al. 2021; Laviola et al. 2022; Madeira et al. 2024; Morales-Marroquín et al. 2025). However, in natural populations of *A. aculeata*, where plants are not part of controlled experiments, spatial models are necessary to account for environmental factors that may influence individual performance during genotype selection. These models help adjust for the overestimation of individuals growing in more favorable microenvironments, resulting in more accurate genotype selection (Ragalzi et al. 2023). For an orphan crop like *A. aculeata*, leveraging its genetic diversity can accelerate the selection of superior genotypes while reducing the drawbacks of mass selection and enhancing breeding efficiency.

Genomic selection (GS) employs genetic marker data to predict the performance of genotypes facilitating population management and the selection process without relying solely on field trials, thus reducing costs and speeding up the identification of promising individuals (Grattapaglia 2022; Díaz-Hernández et al. 2024). GS still allows for the estimation of an individual's genetic value through models that integrate both field and genomic data, making it particularly useful in the absence of detailed pedigree information (Resende 2024). This is particularly relevant for *A. aculeata*, where mixed mating complicates the establishment of complete genealogies. Studies employing genomic approach with SNP markers to identify candidate genes linked to vegetative traits and oil production in *A. aculeata* and GS through genotypic data and optimized training sets as strategies have been conducted (Couto et al. 2024a, b). However, the application of landscape genomics and genomic selection in tropical perennial species like *A. aculeata* remains underdeveloped and faces several challenges.

Collecting and analyzing reliable genetic data, along with addressing the complexities of genotype-environment interactions, requires careful planning (Resende et al. 2021). Predictive maps can be valuable tools for guiding the planting and management of natural populations, helping to identify areas most suitable for specific traits of *A. aculeata*, including those related to its fruit components. Moreover, existing field data can be analyzed more efficiently using spatial models, minimizing the need for controlled experiments. This approach enables more precise selection of genotypes that are well-suited to different environmental conditions, making breeding efforts more effective (Yu et al. 2022; Resende et al. 2020). By integrating spatial and genomic data, these models can optimize population management and enhance the efficiency of breeding programs without relying solely on traditional experimental designs (Lanes et al. 2018; Rico 2021).

In light of the above, this study aimed to investigate genetic variation in *A. aculeata* populations distributed across various landscapes, using genomic and environmental data to identify genotypes adapted to specific local environmental conditions. Additionally, the research sought to integrate

landscape genomics with genomic selection by collecting field data to identify which trees are best suited for seed collection or cloning. This approach could provide a pathway to develop more productive *A. aculeata* cultivars and optimize the use of available genetic resources.

2. Material and Methods

2.1. Plant Materials and Site

In October 2019, two hundred and one trees from natural populations of *A. aculeata* were randomly selected, prioritizing those with mature fruits. The individuals evaluated in this study originate from wild populations. The selected trees are located in Dourado, São Paulo, Brazil (geographic coordinates 22° 6' 13" S, 48° 18' 50" W) and are distributed across rural areas in four provenances with distinct topography and neighboring vegetation (Fig. 1). Provenance P1 comprised 80 individuals, whereas P2, P3, and P4 included 28, 42, and 17 individuals, respectively. Given the natural distribution of the trees, this study did not follow an experimental design. More detailed information on data collection is available in Couto et al. 2024a. The mature fruits were collected from the ground in two different years (year 1 = October 2019 and year 2 = January 2021) and divided into four fruit fractions: husk, pulp, endocarp, and kernel. For this study, data on pulp dry mass (PDM), kernel dry mass (KDM), and pulp oil content (OC) were gathered. The PDM and KDM were measured in grams using a digital scale after drying the pulp and kernel samples in a ventilated oven at 36°C, while OC was determined from the pulp dry mass through Near-Infrared Spectroscopy. This study complies with relevant institutional, national, and international guidelines and legislation. The appropriate permissions for the collection of plant material were taken, and the collections were registered according to the National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SISGEN), as stated by Brazilian Decree No. 8,772 (May 11, 2016) and regulated by Brazilian Law No. 13,123 (May 20, 2015; SISGEN number: A411583, Brazil).

2.2. Genotyping-based methods

Genomic DNA from leaf samples of the 201 selected trees was extracted using the Doyle and Doyle protocol (Doyle and Doyle 1990). Genotyping was conducted using the genotyping-by-sequencing method with the restriction enzymes NsiI and MseI, following the protocol of Poland et al. (2012), with modifications made by Díaz et al. (2021). The final genomic libraries were sequenced in a single run on an Illumina HiSeq3000, configured with single-end and 101bp settings. After the quality control and demultiplexing of the sequencing reads, SNP calling was performed in three different ways, as *A. aculeata* does not have a reference genome available: using the *A. aculeata* transcriptome sequences (Bazzo et al. 2018), the oil palm (*Elaeis guineensis*) reference genome EG5 (NCBI GCA 000442705.1), and the *de novo* pipeline (Stacks v.1.42) (Catchen et al. 2011). SNPs were filtered according to the following criteria: maximum number of alleles = 2, minor allele frequency ≥ 0.01, sequencing depth ≥ 3X, mapping quality ≥ 20, maximum percentage of 30% missing data per locus, and 45% missing data per individual. After, a total of 4,329 SNP was identified in 167 individuals using the transcriptome sequences, 10,444 SNP in 158 individuals using the *Elaeis guineensis* genome, and 27,410 SNP in 153 individuals using the *de novo* pipeline. Missing data were imputed using Beagle 5.3 software (Browning et al. 2021).

2.3. Statistical Modeling on Landscape Genomics

2.3.1. Preparation and Quality control of SNP data

The three genotypic data sets from the SNP-calling step (transcriptome, oil palm genome, and *de novo* pipeline) were initially loaded from VCF files. The vcfR2genind() function from the vcfR package (Knaus and Grünwald 2016, 2017) was used to convert the genotypic data into $\geq n \in d$ objects. Next, the genotypic data were converted to data frames, and the individuals were adjusted to ensure consistency across the different datasets. Of the 201 trees initially evaluated, those present in at least one of the genotypic datasets were retained, leaving 167 trees in this study. The minor allele frequency (MAF) for each SNP was calculated as the proportion of the less frequent allele among the total alleles observed, using the formula $MAF = min(\rho, 1 - \rho)$, where p represents the frequency of the reference allele, with SNPs having a MAF below 0.01 being removed. Subsequently, the processed SNP sets from the genotypic data were combined into a single data frame referred to as a 'combined' dataset (transcriptome + oil palm genome + de novo pipeline). Thus, the genomic matrices used in this study will be referred to as: G_{trs} (transcriptome), G_{den} (oil palm genome), G_{dno} (de novo pipeline), and G_{com} (combined).

2.3.2. Genomic analyses

The four genomic relationship matrices were calculated using the VanRaden (2008) and Vitezica et al. (2013) methods, available in the AGHmatrix package (Amadeu et al. 2023). To assess the similarity between the genomic matrices, pair plots were created using the ggpairs() function from the GGally package (Schloerke et al. 2024). Additionally, a population structure analysis was performed to identify the optimal clusters, employing hierarchical clustering methods with the find.clusters() function from the adegenet package (Jombart 2008; Jombart and Ahmed 2011). The spatial distribution of the clusters was visualized through plots, and a heatmap of the combined genomic matrix was generated.

2.3.3. Phenotypic Analyses and Genetic Effects

The phenotypic data for the traits PDM, KDM, and OC were filtered to retain only the 167 individuals present in the genotypic data from transcriptome, oil palm genome, and *de novo* pipeline. The geographic coordinates of the trees were aggregated, and a geographic distance matrix was calculated using the geosphere package (Hijmans 2024).

Using the phenotypic, genotypic, and geographic distance matrices, we fitted a mixed model for the three traits (PDM, KDM, and OC). This incorporated both genetic and spatial effects, being a GBLUP model extension, employing the mmer() function from the sommer package (Covarrubias-Pazaran 2016). The extended genomic mixed linear model can be expressed with the following notation:

$$y = Xb + Z_1a + Z_2d + Ws + e$$

,

where y is the vector of phenotypic observations, X is the incidence matrix for the fixed effects, b is the vector of fixed effects (in this case, terrain altitude, as shown in Fig. 1), Z_1a, Z_2d, and Ws are the incidence matrices for the random additive (*a*), dominance (*d*), and spatial effects (*s*), respectively, and e is the vector of residual errors. Specifically, Ws models the spatial effects using the function spl2Da(Long, Lat) (adapted from Rodriguez-Alvarez et al. (2018); Lee et al. (2013), Z_1a models the additive genetic effects as vsr(IDg, Gu = G), and Z_2d models the dominance genetic effects as vsr(IDd, Gu = D).

We estimated the variance components and calculated genomic estimated breeding values (GEBV) and genomic expected genetic values (GEGV) for phenotypic traits. The explained genomic variance was interpreted as broad-sense heritability (h^2), while the additive genomic variance was interpreted as narrow-sense heritability (h_a^2). Additionally, an analysis of spatial genetic effects was conducted by fitting a mixed model with both genetic and spatial effects, and the spatial distribution of genetic effects was visualized using Random Forest. To validate the models, the Leave-One-Out (LOO) cross-validation methodology and site-specific validation for the phenotypic traits were performed. The correlations between observed and predicted values were calculated to evaluate of the predictive capacity of the fitted models.

2.3.4. Population Statistics and Genetic Structure

Basic statistics, including observed heterozygosity (Ho), expected heterozygosity (He), and the inbreeding coefficient (F_{IS}), were computed by provenance using the genotypic data from the combined matrix. The data were initially transformed into a $\geq n \in d$ object using the df2genind() function from the adegenet package (Jombart 2008; Jombart and Ahmed 2011), and then converted to the hierfstat format using the genind2hierfstat() function. The basic statistics were calculated using the basic.stats() function from the hierfstat package (Goudet 2005), and specific parameters by provenance were extracted.

To calculate the F_{ST} for each provenance relative to the total (P1 + P2+P3 + P4), we developed a function that selects the specific provenance and the remaining provenances, creating a combined dataframe with adjusted labels. This dataframe was converted into a genind object and subsequently into the hierfstat format, enabling F_{ST} calculation using the wc() function from the hierfstat package (Goudet 2005). Additionally, we conducted a pairwise F_{ST} analysis across all provenance combinations. A custom function was used to calculate pairwise F_{ST} values between two specific provenances, following a procedure similar to that described earlier.

2.4. Summary of the study workflow

To facilitate a comprehensive understanding of the analyses conducted in this study, a flowchart was developed to summarize the main steps. The scheme encompasses processes ranging from SNP preparation and quality control, construction of genomic matrices and exploratory analyses, to the estimation of genetic parameters, predictive validation, and spatial mapping of effects. The steps are organized sequentially and also highlight population statistics and the final outputs, such as heritabilities, genetic structure, and recommended trees (Fig. 2).

3. Results

The phenotypic and genotypic data used in this study were obtained from *A. aculeata* trees distributed across rural properties located near a conservation area, within the Atlantic Forest biome. The study area, the 201 selected trees, and the phenotypic traits evaluated (PDM, KDM, and OC) are presented in Fig. 1. The map in Fig. 1 depicts the spatial distribution of the sampled trees in southeastern Brazil, with an enlarged inset highlighting the four provenances where data were collected (Fig. 1A). The histograms in Fig. 1B display the distribution of PDM, KDM, and OC values assessed over two production years. A substantial overlap between years was observed, indicating no marked variation in trait distributions across the evaluated periods. Additionally, a detailed map of the four *A. aculeata* provenances is provided in Fig. 1, illustrating contour lines that represent the altitudinal gradient of the study area (Fig. 1C). Each provenance was coded with a distinct color, P1 (red), P2 (green), P3 (blue), and P4 (orange), to facilitate visualization of both topographic variation and the spatial arrangement of the 201 sampled trees. The four provenances were separated by approximately 500 meters, with P1 and P2 located at higher elevations compared to P3 and P4.

The characteristics of each *A. aculeata* provenance (P1, P2, P3, and P4), as well as those of the overall population, were evaluated in terms of the number of individuals, phenotypic data collected per year, mean and standard deviation of traits (PDM, KDM, and OC), and genetic diversity parameters (Table 1). Provenance P1 comprised 80 individuals, whereas P2, P3, and P4 included 28, 42, and 17 individuals, respectively. The overall population consisted of 167 individuals, representing those present in at least one of the genotypic datasets used in this study, as described previously in the methodology. The mean PDM values ranged from 19.11 in provenance P3 to 20.54 in P4. For KDM, the means varied from 2.78 in P3 to 3.34 in P2, whereas for OC, the values ranged from 35.35 in P1 to 37.34 in P4. Observed heterozygosity (H_O) and expected heterozygosity (H_E) exhibited identical values across all provenances, as well as in the total population, being 0.96 and 0.55, respectively. Consequently, the inbreeding coefficients (F_{IS}) exhibited negative values. The overall and pairwise F_{ST} values were low, indicating limited genetic differentiation among trees from the different

provenances. Pairwise F_{ST} analysis, considering the total population (P1 + P2+P3 + P4), further supported this low level of differentiation, with values ranging from 0.001 to 0.003.

Table 1
Characteristics of the four *Acrocomia aculeata* provenances (P1, P2, P3, P4) and the overall population (P1 + P2+P3 + P4) for the traits pulp dry mass (PDM), kernel dry mass (KDM), and oil content (OC). The table shows the number of individuals (IDs), the amount of data collected by year (numbers indicate available records for PDM, KDM, and OC, respectively), and trait means (± SD). Genetic diversity parameters, including observed heterozygosity (H_O), expected heterozygosity (H_E), inbreeding coefficient (F_{IS}), and fixation index values (F_{ST}; overall and pairwise between provenances), were calculated using the combined genomic relationship matrix (G_{COM}).

Provenance		**P1**	**P2**	**P3**	**P4**	**Overall**
Number of ID's		80	28	42	17	167
Collected Data by Year	2019 (PDM|KDM|OC)	44 | 44 | 39	28 | 28 | 28	36 | 36 | 33	16 | 16 | 15	124 | 124 | 115
	2021 (PDM|KDM|OC)	70 | 70 | 68	28 | 28 | 28	38 | 38 | 38	17 | 17 | 17	153 | 153 | 151
Trait means ± sd	PDM	19.80 ± 7.00	20.13 ± 5.31	19.11 ± 4.32	20.54 ± 6.82	19.77 ± 6.02
	KDM	3.05 ± 0.76	3.34 ± 0.58	2.78 ± 0.63	3.11 ± 0.54	3.04 ± 0.69
	OC	35.35 ± 8.24	36.23 ± 6.91	37.09 ± 6.27	37.34 ± 5.13	36.24 ± 7.16
H_O		0.96	0.96	0.96	0.96	0.96
H_E		0.55	0.55	0.55	0.55	0.55
F_{IS}		-0.75	-0.75	-0.76	-0.76	-0.74
		P1	**P2**	**P3**	**P4**	**Overall**
F_{ST} (paired)	**P1**	–	–	–	–	0.002
	P2	0.003	–	–	–	0.003
	P3	0.003	0.004	–	–	0.002
	P4	0.002	0.002	0.001	–	0.001

To investigate the relationships among the different genomic matrices, G_{trs} (Transcriptome), G_{den} (Oil Palm Genome), G_{dno} (de novo) e G_{com} (Combined) correlation analyses were performed. In addition, a heatmap of the combined genomic matrix was generated to assess the genetic similarity among individuals (Fig. 3). The relationships among the four genomic matrices are presented in Fig. 3A. The correlations among G_{trs}, G_{den}, G_{dno} with G_{com} were 0.904, 0.842, and 0.992, respectively. The lowest correlation (0.482) was observed between G_{trs}, and G_{den}, whereas the highest (0.989) occurred between G_{com} and G_{dno}. The heatmap of the combined genomic matrix depicting genetic similarity among the 167 individuals is shown in Fig. 3B. The dendrogram on the left represents the hierarchical clustering based on genetic similarity, while the color bars on the right indicate the provenance of each individual.

Genomic similarity analysis of the total population was conducted to investigate the presence of subpopulations, and the clustering of trees into different subpopulations is presented in Fig. 4. In the upper-left corner of the figure, a clear decline in genomic similarity with increasing geographic distance is observed. The upper-right panel shows that the optimal number of clusters determined by Ward's method was *K* = 2. To further explore the subpopulation structure and its relationship with genomic similarity, we also considered *K* = 4, corresponding to the number of provenances studied. For *K* = 2, the provenances appear to belong to a single population, with divergent individuals represented in red and blue. When *K* = 4, greater genetic diversity is observed within provenance P1, whereas P2, P3, and P4 exhibit higher genetic similarity.

From the statistical GBLUP model, different variance components were estimated, along with the proportion of variance explained for each evaluated trait, using the following SNP calling approaches: transcriptome, oil palm genome, *de novo* pipeline, and combined (Table 2). Substantial variation was detected across the different SNP-calling pipelines. Among the variance components, the PDM trait exhibited genomic additivity values ranging from 15.07 in the *de novo* pipeline to 36.41 in the transcriptome SNP calling. Genomic dominance values for PDM ranged from 0.00 to 7.50, while the spatial kernel varied from 17.66 in the combined matrix to 21.03 in the de novo pipeline. For KDM, genomic additivity values ranged from 0.17 in the *de novo* pipeline to 0.51 in the transcriptome SNP calling. Genomic dominance values for KDM ranged from 0.01 in the oil palm genome to 0.09 in the transcriptome SNP calling, while spatial kernel values varied between 0.12 and 0.28. For OC, the highest genomic additivity (42.85) and genomic dominance (29.27) values were observed in the transcriptome SNP calling, whereas the lowest values were obtained in the de novo pipeline (18.71 and 9.35, respectively). Spatial kernel values for OC ranged from 61.42 to 100.16. The explained variance (%) results presented in Table 2 represent the

proportion of the observed values in the variance components for each trait and SNP calling. The genomic additive component was higher than the genomic dominance component across all SNP callings and traits and, in certain cases, the spatial kernel value surpassed the total genomic variance

The predictive abilities of genomic selection were evaluated for PDM, KDM, and OC across different SNP calling under two cross-validation schemes. This analysis aimed to assess the efficiency of the genomic models in predicting individual genetic values, both from an additive perspective (GEBV) and a total genetic perspective (GEGV), and to examine the consistency of predictions across distinct genomic datasets and provenances. The results showed that predictive abilities for all traits varied depending on the SNP-calling and the validation scheme applied (Table 3). Overall, the predictions under the LOO validation scheme were slightly higher than those observed in cross-validation by provenance, reflecting greater accuracy when the model is evaluated within its own training population. Among the different SNP callings, the oil palm genome and combined genomic matrix exhibited the highest predictive abilities for most traits, indicating that increased genomic coverage and the integration of multiple SNP origins contributed positively to selection accuracy. The GEBV and GEGV values were similar across all scenarios, both under the LOO and provenance validation schemes, suggesting that additive effects explain the majority of the genetic variation for the evaluated traits. For PDM and KDM, the highest predictive abilities in the LOO validation scheme were observed when using the oil palm genome and combined genomic matrix, for both GEBV and GEGV. For OC, prediction values were the highest among all traits studied, reaching correlations of up to 0.50 (GEGV, transcriptome) and 0.48 (GEGV and GEBV, oil palm genome), indicating that this trait was the most responsive to genomic selection. The interpretation of results by trait and SNP calling for the cross-validation by provenance was similar to that observed under the LOO scheme. This trend is consistent with the predominance of additive variance observed in Table 2, suggesting that the inclusion of dominance effects did not lead to substantial gains in prediction accuracy in our work. Therefore, the additive model proved sufficient to capture the genetic variation associated with PDM, KDM, and OC in the studied population.

Table 2
Modeled variance components and explained variance for pulp dry mass (PDM), kernel dry mass (KDM), and oil content (OC) using different SNP calling strategies: transcriptome, oil palm genome, *de novo* pipeline, and combined matrix. The table shows the variance components and explained variance calculated for each trait, including spatial kernel, genomic additive variance (Genomic Additivity), genomic dominance variance (Genomic Dominance), and residual variance. The explained variance is presented as a percentage for each component.

Trait	SNP Source	Variance Components				Explained Variance (%)					
		Spatial Kernel	Genomic Additivity	Genomic Dominance	Residual	Spatial Kernel	Genomic Total*	Genomic Additivity**	Genomic Dominance	Total	Residual
PDM	Transcriptome	19.67	36.41	7.50	11.81	26.09%	58.25%	48.30%	9.95%	84.34%	15.66%
	Oil palm genome	18.41	27.72	0.00	13.43	30.91%	46.54%	46.54%	0.00%	77.45%	22.55%
	de novo pipeline	21.03	15.07	0.00	13.67	42.26%	30.29%	30.29%	0.00%	72.54%	27.46%
	Combined	17.66	21.52	0.00	12.72	34.02%	41.47%	41.47%	0.00%	75.49%	24.51%
KDM	Transcriptome	0.12	0.51	0.09	0.16	13.31%	68.80%	58.57%	10.23%	82.11%	17.89%
	Oil palm genome	0.28	0.32	0.01	0.19	35.41%	40.71%	39.88%	0.82%	76.12%	23.88%
	de novo pipeline	0.28	0.17	0.03	0.20	41.32%	28.83%	24.81%	4.02%	70.15%	29.85%
	Combined	0.26	0.25	0.03	0.18	35.97%	39.00%	34.84%	4.16%	74.98%	25.02%
OC	Transcriptome	61.42	42.86	29.28	13.84	41.67%	48.94%	29.08%	19.86%	90.61%	9.39%
	Oil palm genome	83.89	33.45	13.25	14.89	57.67%	32.10%	22.99%	9.11%	89.77%	10.23%
	de novo pipeline	100.16	18.71	9.35	16.26	69.32%	19.43%	12.95%	6.47%	88.74%	11.26%
	Combined	75.33	23.65	15.69	14.59	58.28%	30.44%	18.30%	12.14%	88.72%	11.28%

* The explained total genomic variance, in this case, can also be understood as broad-sense heritability (h^2); and ** the additive genomic variance as narrow-sense heritability (h^2_A).

Table 3

Predictive abilities of Genomic Selection (GS) for pulp dry mass (PDM), kernel dry mass (KDM), and oil content (OC) using different SNP calling. The table presents the predictive abilities measured as genomic estimated breeding values (GEBV) and genomic estimated genetic values (GEGV) under two validation schemes: Leave-One-Out (LOO) and Cross-Validation by Provenance. Results are shown for three SNP calling: transcriptome, oil palm genome, and *de novo* pipeline, as well as a fourth matrix combining all three sources. The GEBV represents the additive genetic effects, while the GEGV includes both additive and dominant genetic effects. Cross-validation results are presented as means ± standard deviations.

Trait	**SNP Source***	**Leave One Out (LOO)**		**Cross-Validation by Provenance**	
		GEBV (additive)	**GEGV (total genetics)**	**GEBV (additive)**	**GEGV (total genetics)**
PDM	Transcriptome	0.25	0.25	0.14 ± 0.32	0.16 ± 0.34
	Oil palm genome	0.35	0.35	0.28 ± 0.30	0.26 ± 0.3
	de novo pipeline	0.34	0.34	0.27 ± 0.27	0.27 ± 0.27
	Combined	0.35	0.35	0.27 ± 0.29	0.27 ± 0.29
KDM	Transcriptome	0.30	0.28	0.03 ± 0.26	0.02 ± 0.27
	Oil Palm genome	0.43	0.43	0.10 ± 0.22	0.10 ± 0.23
	de novo pipeline	0.39	0.39	0.18 ± 0.25	0.18 ± 0.25
	Combined	0.40	0.40	0.16 ± 0.25	0.16 ± 0.25
OC	Transcriptome	0.44	0.50	0.21 ± 0.15	0.33 ± 0.09
	Oil palm genome	0.48	0.48	0.25 ± 0.09	0.25 ± 0.09
	de novo pipeline	0.45	0.44	0.17 ± 0.15	0.17 ± 0.15
	Combined	0.47	0.46	0.19 ± 0.14	0.19 ± 0.14

*Using both the genomic additive matrix (G) and the genomic dominance matrix (D). Total genetics = $G + D$

The results of predictive abilities, obtained by validating genomic selection models across distinct provenances to assess their transferability between different genetic backgrounds and SNP sources, are presented in Table 4. Predictive abilities varied according to the traits, the SNP-calling used (including the combined genomic matrix), and the provenance employed for validation, demonstrating a moderate to high potential for the application of genomic selection across *A. aculeata* provenances. Only predictive abilities for GEBV (additive effects) above the overall mean (0.18) are shown, with values that were not achieved or did not converge being omitted. For PDM, the highest predictive abilities were obtained when the model was trained on provenances 1, 2, and 3 and validated on provenance 4, reaching 0.58 for both the oil palm genome and the combined genomic matrix, indicating good transferability of models across genetic backgrounds. In general, GEGV values followed a similar trend to those of GEBV. For KDM, predictive abilities were more consistent but slightly lower than those observed for PDM, ranging from 0.29 to 0.35. The best result for this trait was obtained using the *de novo* pipeline, training on provenances 1, 2, and 3 and validating on provenance 4 (0.35). For OC, predictive ability values ranged from 0.19 to 0.40, with the highest (0.40) achieved using the transcriptome dataset trained on provenances 1, 3, and 4 and validating on provenance 2. Across traits, combined SNP datasets generally yielded the most accurate and stable predictions. These findings emphasize that both the composition of the training population and the SNP source are critical for maximizing cross-provenance prediction accuracy in *A. aculeata*, particularly when aiming to develop broadly applicable genomic selection models. These findings emphasize that both the composition of the training population and the SNP source are critical for maximizing cross-provenance prediction accuracy in *A. aculeata*, particularly when aiming to develop broadly applicable genomic selection models.

Table 4
Validated predictive abilities for each provenance (training the GS model on the remaining three provenances) for pulp dry mass (PDM), kernel dry mass (KDM), and oil content (OC) using different SNP callings. Only predictive abilities for GEBV above the average (> 0.189) are shown. Values that were not achieved (or did not converge) are also omitted.

Trait	SNP Calling*	Training Set (provenances)	Validation Set (provenances)	GEBV (additive)	GEGV (total genetics)
PDM	Transcriptome	1,2,3 →	4	0.50	0.54
	Oil palm genome	1,3,4 →	2	0.30	0.22
	Oil palm genome	1,2,3 →	4	0.58	0.58
	de novo pipeline	1,3,4 →	2	0.22	0.22
	de novo pipeline	1,2,3 →	4	0.56	0.56
	Combined	1,3,4 →	2	0.24	0.24
	Combined	1,2,3 →	4	0.58	0.58
KDM	Transcriptome	1,2,4 →	3	0.31	0.31
	Oil palm genome	1,2,4 →	3	0.30	0.30
	de novo pipeline	1,2,4 →	3	0.29	0.29
	de novo pipeline	1,2,3 →	4	0.35	0.35
	Combined	1,2,4 →	3	0.30	0.30
	Combined	1,2,3 →	4	0.30	0.30
OC	Transcriptome	1,3,4 →	2	0.32	0.40
	Oil Palm Genome	1,3,4 →	2	0.31	0.31
	Oil Palm Genome	1,2,3 →	4	0.19	0.19
	de novo pipeline	1,3,4 →	2	0.28	0.28
	Combined	1,3,4 →	2	0.29	0.29

*Using both the genomic additive matrix (G) and the genomic dominance matrix (D). Total genetics = $G + D$

The spatial analysis of predicted genetic effects enables visualization of the geographical distribution of populations and the identification of superior individuals for use in breeding programs and seed collection. This approach enables the assessment of potential spatial patterns associated with genetic effects, thereby optimizing strategies for the conservation and utilization of genetic resources. Based on this premise, the predicted genetic effects for *A. aculeata* were interpolated using Random Forest models derived from the combined genomic matrix (G_{com}). Figure 5 presents the results of this interpolation for PDM, KDM, and OC. For all traits, areas with varying levels of genetic effects were observed where cooler colors represent higher genetic values, highlighting spatial variation in the genetic potential of the studied individuals. The genotypes showing high genetic values for each evaluated trait are highlighted in purple. For PDM, the standout genotypes were EC77 and EC201. For KDM, genotype EC49 showed the highest performance, while for OC, EC43 and EC44 were the most prominent. Some genotypes, such as EC102 and EC165, stood out among those with the highest predicted genetic effects, consistently appearing for both PDM and OC, as well as for PDM and KDM, suggesting a positive genetic correlation among these traits. However, areas with higher genetic effects did not exhibit distinct spatial patterns, indicating that high-performing genotypes are not clustered in specific regions. The results show that areas with higher effects are not spatially clustered, indicating that seed collection strategies should prioritize individual trees with superior genetic potential distributed across different locations, rather than focusing on particular geographic areas.

4. Discussion

4.1. Landscape genomics across the four geographic provenances of *A. aculeata*

The four provenances evaluated exhibited strong genetic diversity within populations and minimal differentiation among them (H = 0.96). The expected heterozygosity (H) remained constant among provenances at 0.55, resulting in negative inbreeding coefficients ($F_{IS} \cong$ —-0.75). These results are consistent with a predominantly outcrossing, and likely self-incompatible reproductive system reported for *A. aculeata* (Laviola et al. 2022). Similar patterns were previously observed in studies using both SNP markers (Couto et al. 2024a) and microsatellites (Lanes et al. 2016; de Lima et al. 2020), with genetic differentiation also influenced by the caching behavior of small mammal seed dispersers (Mengistu et al. 2016; de Lima et al. 2020; Laviola et al. 2022).

Global genetic differentiation (F_{ST}) values were extremely low, ranging from 0.001 to 0.003, indicating that more than 99.7% of the genotypic variation occurs within provenances. Pairwise values confirmed this pattern, ranging from 0.001 (P3×P4) to 0.004 (P2×P3). These results suggest the absence of major physical barriers on this geographic scale (Fig. 1). Although exhibiting differences in elevation and microhabitat, the provenances are spatially close, facilitating pollen and seed dispersal by medium and long-distance pollinators and dispersers (Araújo et al. 2017; Resende et al. 2020). This pattern is characteristic of widely distributed tropical species that, in the absence of severe habitat fragmentation, exhibit weak neutral genetic structure (Holderegger and Wagner 2006; Lanes et al. 2018). It is worth noting that the diversity estimates (Table 1) were obtained using the combined genomic relationship matrix (G_{com}), which integrates SNPs from three SNP-calling pipelines. This approach showed higher heterozygosity estimates than those reported by Couto et al. (2024a), reflecting the broader genomic coverage and the high correlation among matrices observed (Fig. 3A).

Genomic similarity analyses (Fig. 3) and spatial clustering results (Fig. 4) corroborated the connectivity observed among provenances. The relationship between genomic similarity and geographic distance indicated the absence of isolation by distance, and the clusters defined at K = 2 or K = 4 did not correspond to any significant environmental barriers. Thus, the individuals analyzed belong to a single population (P1 + P2 + P3 + P4 = one population). The optimal number of clusters determined by Ward's method was K = 2, revealing two genetically distinct groups within the population. The cluster represented by the red circle was dominant, as it occurred at higher frequencies across all provenances, whereas the cluster represented by the blue triangle was distributed across P1, P2, and P3. Although K = 4 reflected the four provenances highlighted in this study, this result should be interpreted with caution, as structural analyses do not allow inferences about ancestral sources. Nevertheless, the results suggest that provenance P1 likely represents the ancestral source of the others, as previously reported by Couto et al. (2024a). Given that P1 is located at a higher altitude and that *Acrocomia aculeata* fruits are rounded, their shape likely facilitated seed dispersal toward lower-altitude regions.

From a conservation perspective, the combination of high H, negative F_{IS}, and low F_{ST} observed in our results is encouraging, as it indicates a broad genetic base and strong population connectivity. These attributes enhance the adaptive potential of *A. aculeata* populations in the face of environmental change and reduce the risk of inbreeding depression (Kremer and Yeaman 2017). However, as highlighted by Lima et al. (2020), habitat fragmentation or declines in pollinator abundance could rapidly reverse this scenario, leading to increased inbreeding and loss of genetic variability. Recent studies, including plastome analysis, have revealed that *A. aculeata* has undergone positive selection and distinct genetic clustering, influenced by biogeographic barriers and dispersal processes, with selective signatures linked to key metabolic and physiological processes, pathogen resistance, and adaptation to environmental stress (de Santana Lopes et al. 2018; Morales-Marroquín et al. 2025). Therefore, conservation and germplasm collection programs should ensure sampling individual trees across multiple provenances and maintain the integrity of gene flow corridors (Laviola et al., 2022; Díaz-Hernández et al., 2024).

4.2. Genomic prediction models showed satisfactory predictive abilities when applied to *A. aculeata* trees.

The decomposition of variance components (Table 2) showed that the mixed models incorporating both spatial and genomic information captured a substantial proportion of the phenotypic variation for PDM, KDM, and OC. The relative contribution of the spatial kernel ranged from 13.31% (KDM, transcriptome) to 69.32% (OC, *de novo* pipeline), indicating that microenvironmental factors exert variable influence depending on the trait and SNP datasets (Resende et al., 2020; Ragalzi et al., 2023). Genomic additive variance was particularly high for OC, reaching 42.86 (transcriptome SNP calling) and 33.45 (oil palm genome), suggesting strong genetic control of this trait, a pattern consistent with estimates reported for oil palm (*Elaeis guineensis*) in genomic prediction models (Kwong et al., 2017). Dominance variance, although generally lower, was relevant in specific scenarios, particularly for PDM and KDM, confirming that non-additive effects should also be considered in the modeling process (Bekele et al., 2022; Vitezica et al., 2013).

The predictive accuracies under the LOO validation scheme were consistent with this pattern (Table 3), with GEBV and GEGV reaching low to medium values. In turn, the inclusion of dominance effects did not lead to substantial gains in prediction accuracy in our work. Therefore, the additive model proved sufficient to capture the genetic variation associated with PDM, KDM, and OC in the studied population. The cross-validation by provenance scheme, which is more stringent as it tests the ability of models to generalize across geographic groups, accuracies decreased but remained satisfactory for PDM (up to 0.28, GEBV, oil palm genome) and OC (0.33, GEGV, transcriptome). This pattern is expected in populations with wide genetic diversity and low genetic structure, as observed for *A. aculeata* in this study (Lanes et al., 2018; Díaz et al., 2021). In addition, considering that *A. aculeata* is a non-domesticated species and the evaluations were conducted under natural conditions, these values are promising. The provenance-specific cross-validation (Table 4) reinforces this interpretation. The values obtained are compatible with practical applications in genomic selection for perennial plants (Grattapaglia, 2022; Kwong et al., 2017), indicating that even when one provenance is excluded from the training set, substantial predictive transferability remains, likely due to the gene flow and genetic cohesion described in Section 4.1 (Araújo et al., 2017).

The comparison among SNP matrices further shows that the choice of genomic source affects performance: the combined and transcriptome matrices were the most consistent across traits and validation schemes. The first chromosome-scale genome of *Acrocomia aculeata* has been recently obtained; however, the reference genome is not yet publicly available (Scaketti et al., 2025). In this context, strategies that integrate multiple SNP calling sources tend to maximize genomic representativeness and improve accuracy, as reported for *A. aculeata* by Couto et al. (2024b). In general, our findings emphasize that both the composition of the training population and the SNP source are critical for maximizing cross-provenance prediction accuracy in *A. aculeata*, particularly when aiming to develop broadly applicable genomic selection models.

4.3. Defining Seed collection trees based on *A. aculeata* fruit traits

The spatial distribution of predicted genetic effects for PDM, KDM, and OC (Fig. 5) revealed important patterns for defining seed collection strategies. Although some individuals are promising across multiple traits, the genetic basis of each characteristic is relatively specific, a common scenario in outcrossing perennial species, where genetic correlations among yield-related traits are often moderate to low (Kwong et al., 2017; Bekele et al., 2022).

The absence of genotypes with outstanding performance for PDM, KDM, and OC simultaneously suggests that the physiological and genetic processes controlling these traits are largely independent, reinforcing the need for multi-trait selection strategies. This approach is particularly important in *A. aculeata*, as increases in one fruit component do not necessarily result in proportional gains in the others (Couto et al., 2024; Laviola et al., 2022). Therefore, to maximize genetic progress, it is necessary to balance the selection of individuals with superior performance in specific traits with those showing satisfactory performance across multiple traits, as recommended for breeding programs of perennial crops targeting multiple end uses (Grattapaglia, 2022).

The interpolated spatial analysis (Fig. 5) did not reveal clear geographic patterns that would justify seed collection focused on specific areas. This finding is consistent with the high genetic connectivity among provenances described in Section 4.1 and with the low F_{ST} values observed (Table 1), which indicate homogeneous gene flow and a lack of pronounced spatial structure (Díaz et al., 2021; Araújo et al., 2017). Thus, the most robust recommendation is to target seed collection directly from the individual trees identified as genetically superior for the traits of interest, regardless of their spatial location. Adopting this selective strategy offers multiple advantages: (i) it ensures capture of the maximum genetic potential detected by the prediction models (Tables 3 and 4), (ii) it maintains within-population genetic diversity by incorporating individuals from different provenances, and (iii) it enables the establishment of base populations for breeding programs tailored to distinct industrial applications, such as oil production, biofuels, or food uses (Resende et al., 2020; Díaz et al., 2021).

5. Final Considerations

This study demonstrated that the natural populations of *Acrocomia aculeata* evaluated here exhibit high within-population genetic diversity, low differentiation among provenances, and robust genomic connectivity. These results indicate that, at the regional scale, the species forms a genetic continuum in which gene exchange among individuals is extensive and ongoing. This characteristic broadens the opportunities for breeding programs, enabling the selection of promising genotypes across different provenances without significant losses of genetic diversity.

The genomic prediction analyses showed that, even under different SNP-calling strategies and cross-validation schemes, it is possible to achieve satisfactory accuracies for agronomically relevant traits such as pulp dry mass, kernel dry mass, and oil content. This demonstrates that genomic approaches can be efficiently integrated into selection strategies, thereby optimizing genetic gain per cycle.

The identification of individual trees with superior genetic values, regardless of their spatial location, reinforces the importance of selective seed collection strategies focused on high-performing individuals. This approach maximizes the use of available genetic potential and supports the development of more productive and better-adapted plant materials.

In summary, the findings point to a favorable scenario for the conservation and breeding of *A. aculeata*. Maintaining genetic connectivity, investing in genomic prediction tools, and adopting multi-trait selection criteria are strategic steps toward ensuring the long-term sustainability and advancement of the species' cultivation.

Declarations

Acknowledgments

We thank Bárbara Regina Bazzo, Lucas Miguel de Carvalho, and Marcelo Falsarella Carazzolle for providing the *A. aculeata* transcriptome data used in the SNP calling procedures (doi:10.1186/s12870-018-1509-9). We also express our gratitude to Jonathan A. Morales-Marroquín, Carlos Augusto Colombo, Joaquim Adelino de Azevedo-Filho, and other researchers from the Campinas Agronomic Institute for their assistance with phenotypic data collection in the field.

Data availability statement

Publicly available dataset were analyzed in this study. The database in which the DNA data is stored is here: 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 used in the current study are also there.

Author contributions

EGOC: Sample collected, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. LLSP: Formal analysis, Investigation, Writing – review & editing. MSV: Investigation, Writing – original draft, Writing –review & editing. KNK: Investigation, Writing – original draft, Writing –review & editing. DIR: Investigation, Writing – original draft, Writing – review & editing. SYM: Project administration, Funding acquisition, Writing – review & editing. MIZ: Project administration, Funding acquisition, Writing – review & editing. RTR: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing –review & editing. All authors read the manuscript.

Competing interests

The authors declare no competing interests.

Funding Declaration

This study was supported by the São Paulo Research Foundation, FAPESP (grant 2019/20307-0 and Thematic Project grant 14/23591-7); the Financiadora de Estudos e Projetos, FINEP (AcroAlliance Project, Convênio 01.21.0177.00, Ref. No. 0566/21), and the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES).

References

1. Araújo MRG, et al (2017) Fine-scale spatial genetic structure and gene flow in *Acrocomia aculeata* (Arecaceae). Biochem Syst Ecol 71:147–154
2. Amadeu RR, Garcia AAF, Munoz PR, Ferrão LFV (2023) AGHmatrix: genetic relationship matrices in R. Bioinformatics 39. https://doi.org/10.1093/bioinformatics/btad445
3. Bazzo BR, de Carvalho LM, Carazzolle MF, Pereira GAG, Colombo CA (2018) Development of novel EST-SSR markers in the macauba palm (*Acrocomia aculeata*) using transcriptome sequencing and cross-species transferability in Arecaceae species. BMC Plant Biol 18:224. https://doi.org/10.1186/s12870-018-1509-9
4. Bekele F, Bidaisee G, Allegre M, Argout X, Fouet O, Boccara M, Saravanakumar D, Bekele I, Lanaud C (2022) Genome-wide association studies and genomic selection assays made in a large sample of cacao (*Theobroma cacao* L.) germplasm reveal significant marker-trait associations and good predictive value for improving yield potential. PLoS ONE 17:e0260907. https://doi.org/10.1371/journal.pone.0260907
5. Browning BL, Tian X, Zhou Y, Browning SR (2021) Fast two-stage phasing of large-scale sequence data. Am J Hum Genet 108:1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005
6. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3 Genes Genomes Genet 1:171–182. https://doi.org/10.1534/g3.111.000240
7. Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE 11:e0156744. https://doi.org/10.1371/journal.pone.0156744
8. Couto EGO, et al (2024a) Genome-wide association insights into the genomic regions controlling vegetative and oil production traits in *Acrocomia aculeata*. BMC Plant Biol 24:1125. https://doi.org/10.1186/s12870-024-05805-y
9. Couto EGO, et al (2024b) Training set optimization is a feasible alternative for perennial orphan crop domestication and germplasm management: an *Acrocomia aculeata* example. Front Plant Sci 15:1383329. https://doi.org/10.3389/fpls.2024.1441683
10. Díaz BG, et al (2021) Genome-wide SNP analysis to assess the genetic population structure and diversity of *Acrocomia* species. PLoS ONE 16:e0246231. https://doi.org/10.1371/journal.pone.0241025
11. Díaz-Hernández BG, et al (2024) Assessing the genetic vulnerability of macauba palm (*Acrocomia aculeata*) through the mating system and genetic diversity of open-pollinated progenies. Ann Appl Biol 184:238–249. https://doi.org/10.1111/aab.12879
12. Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15
13. Goudet J (2005) hierfstat, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes 5:184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x
14. Grattapaglia D (2022) Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. Forests 13:1554. https://doi.org/10.3390/f13101554
15. Hijmans R (2024) geosphere: spherical trigonometry. R package version 1.5-20. https://github.com/rspatial/geosphere
16. Holderegger R, Wagner HH (2006) A brief guide to landscape genetics. Landscape Ecol 21:793–796. https://doi.org/10.1007/s10980-005-6058-6
17. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. https://doi.org/10.1093/bioinformatics/btn129
18. Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070–3071. https://doi.org/10.1093/bioinformatics/btr521
19. Knaus BJ, Grünwald NJ (2017) vcfR: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour 17:44–53. https://doi.org/10.1111/1755-0998.12549
20. Konzen ER, Zucchi MI (2020) Landscape genetics: from classic molecular markers to genomics. In: Methods in Molecular Medicine. IntechOpen
21. Lanes ECM, Motoike SY, Kuki KN, Resende MDV, Caixeta ET (2016) Mating system and genetic composition of the macaw palm (*Acrocomia aculeata*): implications for breeding and genetic conservation programs. J Hered 107:527–536. https://doi.org/10.1093/jhered/esw038
22. Lanes EC, et al (2018) Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from Amazonian savannas. Front Plant Sci 9:532. https://doi.org/10.3389/fpls.2018.00532
23. Lee DJ, Durban M, Eilers PHC (2013) Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. Comput Stat Data Anal 61:22–37. https://doi.org/10.1016/j.csda.2012.11.013

24. Lima NE, Carvalho AA, Meerow AW, Manfrin MH (2018) A review of the palm genus *Acrocomia*: neotropical green gold. Org Divers Evol 18:151–161. https://doi.org/10.1007/s13127-018-0362-x
25. Lorenzi H, Noblick LR, Kahn F, Ferreira E (2010) Flora brasileira: Arecaceae (palmeiras). Instituto Plantarum
26. Madeira DDC, et al (2024) Phenotypic characterization and genetic diversity of macauba (*Acrocomia aculeata*) accessions based on oil attributes and fruit biometrics. Genet Resour Crop Evol 71:3163–3178. https://doi.org/10.1007/s10722-024-01856-0
27. Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18:189–197
28. Morales-Marroquín JA, et al (2025) Genetic variations associated with adaptation in *Acrocomia* palms: a comparative study across the Neotropics for crop improvement. PLoS ONE 20:e0324340. https://doi.org/10.1371/journal.pone.0324340
29. Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253. https://doi.org/10.1371/journal.pone.0032253
30. Ragalzi CM, et al (2023) A spatial-based approach applied to early selection stages in a forage breeding program. Euphytica 219:58. https://doi.org/10.1007/s10681-023-03179-0
31. Resende RT (2024) Balancing genomic selection efforts for allogamous plant breeding programs. J Crop Sci Biotechnol 27:349–358. https://doi.org/10.1007/s12892-024-00234-7
32. Resende RT, et al (2020) Data-based agroecological zoning of *Acrocomia aculeata*: GIS modeling and ecophysiological aspects into a Brazilian representative occurrence area. Ind Crops Prod 154:112749. https://doi.org/10.1016/j.indcrop.2020.112749
33. Resende RT, et al (2021) Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor Appl Genet 134:95–112. https://doi.org/10.1007/s00122-020-03684-z
34. Rodriguez-Alvarez MX, Boer MP, van Eeuwijk FA, Eilers PHC (2018) SpATS: spatial analysis of field trials with splines. R package version 1.0-9. https://CRAN.R-project.org/package=SpATS
35. Scariot A, Lleras E, Hay JD (1995) Flowering and fruiting phenologies of the palm *Acrocomia aculeata*: patterns and consequences. Biotropica 27:168–173
36. Schloerke B, et al (2024) GGally: extension to ggplot2. R package version 2.2.1. https://ggobi.github.io/ggally
37. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980
38. Vargas-Carpintero R, et al (2021) *Acrocomia* spp.: neglected crop, ballyhooed multipurpose palm or fit for the bioeconomy? Agron Sustain Dev 41:75. https://doi.org/10.1007/s13593-021-00729-5
39. Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223–1230. https://doi.org/10.1534/genetics.113.155176
40. Yu Y, Aitken SN, Rieseberg LH, Wang T (2022) Using landscape genomics to delineate seed and breeding zones for lodgepole pine. New Phytol 235:1653–1664. https://doi.org/10.1111/nph.18228
41. Yaqoob H, et al (2023) Integrating genomics and genome editing for orphan crop improvement: a bridge between orphan crops and modern agriculture system. GM Crops Food 14:1–20. https://doi.org/10.1080/21645698.2022.2146952

Figures

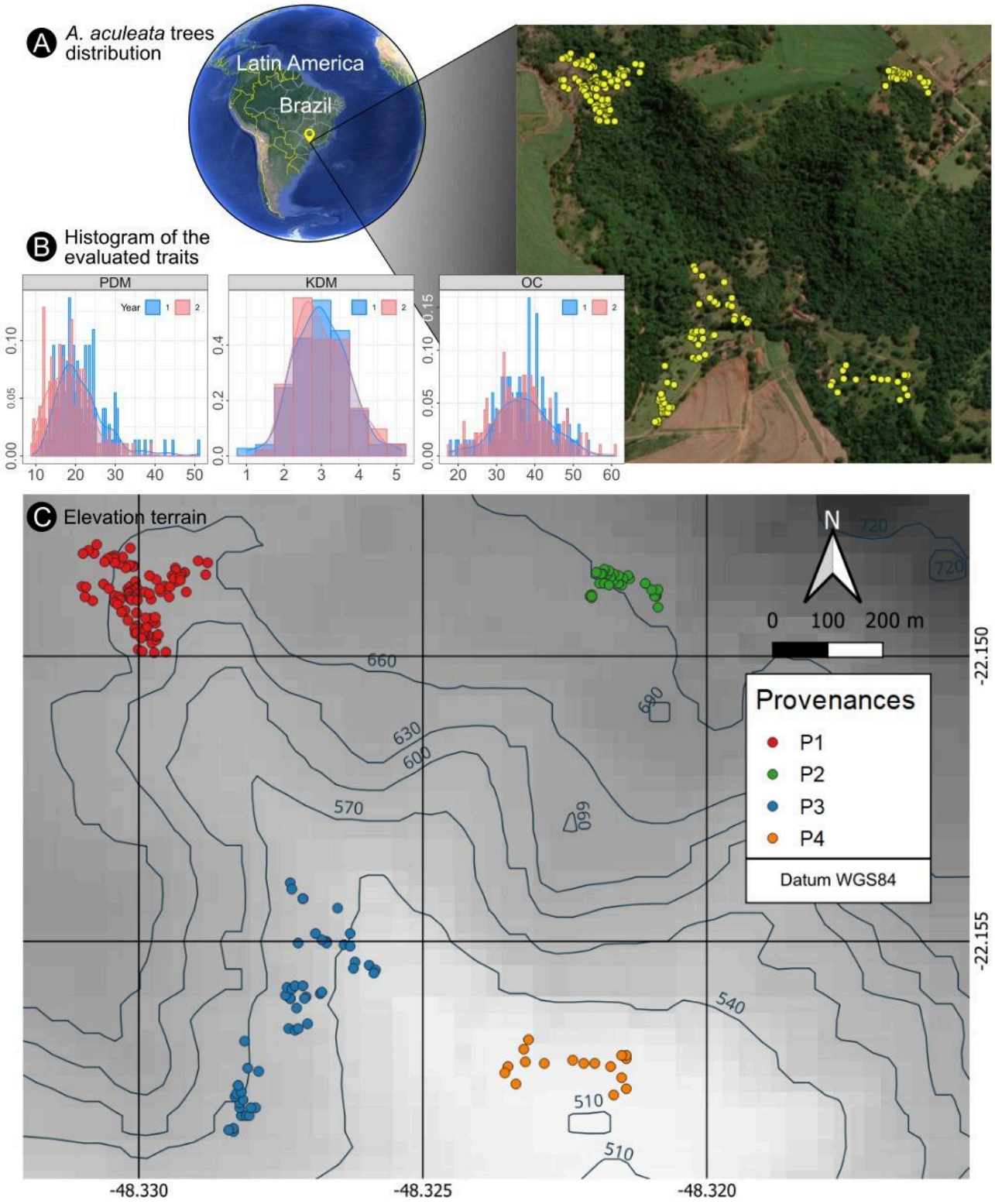

Figure 1

Study area and data collection overview. (A) The map shows the distribution of *Acrocomia aculeata* trees in Dourado, São Paulo, southeastern Brazil. The zoomed-in view highlights the specific area where data were collected, with trees from four different provenances marked. (B) Histograms of the evaluated traits: pulp dry mass (PDM), kernel dry mass (KDM), and oil content (OC), assessed in two years (Year 1 = October 2019, Year 2 = January 2021). (C) Detailed map of the four provenances of *A. aculeata*, shown with elevation contours (altimetric elevation) to illustrate the terrain of the study area. Provenances are color-coded as P1 (red), P2 (green), P3 (blue), and P4 (orange).

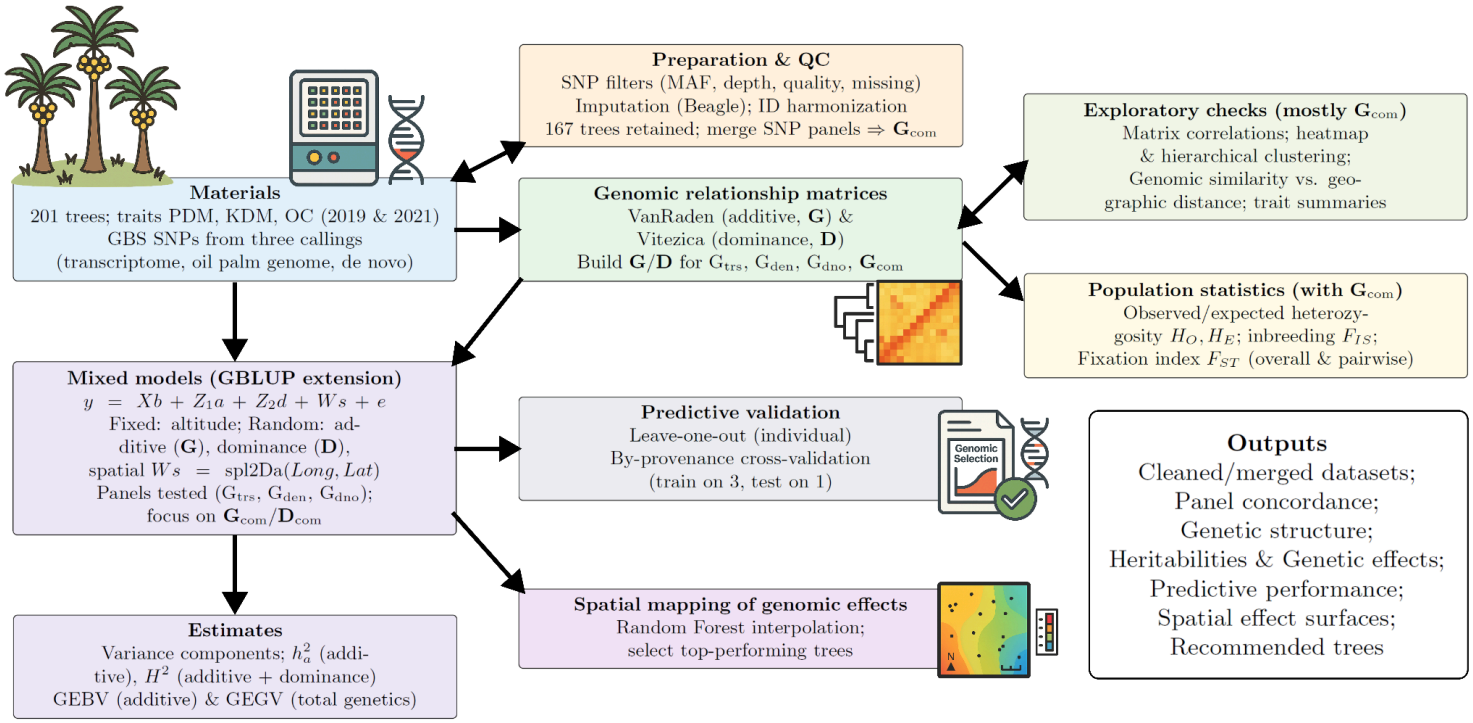

Figure 2

Flowchart of the data analysis pipeline. From phenotypic and genotypic data (GBS SNPs), quality control, imputation, and harmonization led to the construction of genomic relationship matrices. Exploratory analyses and population statistics were mainly based on the combined SNP panel (G_{com}). Mixed models (GBLUP extensions) were fitted to estimate variance components, heritabilities, and genomic values. Predictive ability was assessed by individual and by-provenance cross-validation. Spatial mapping of genomic effects was obtained via Random Forest interpolation. Outputs include genetic structure, population parameters, predictive performance, spatial surfaces of effects, and selection of top-performing trees. (Image generated by artificial intelligence).

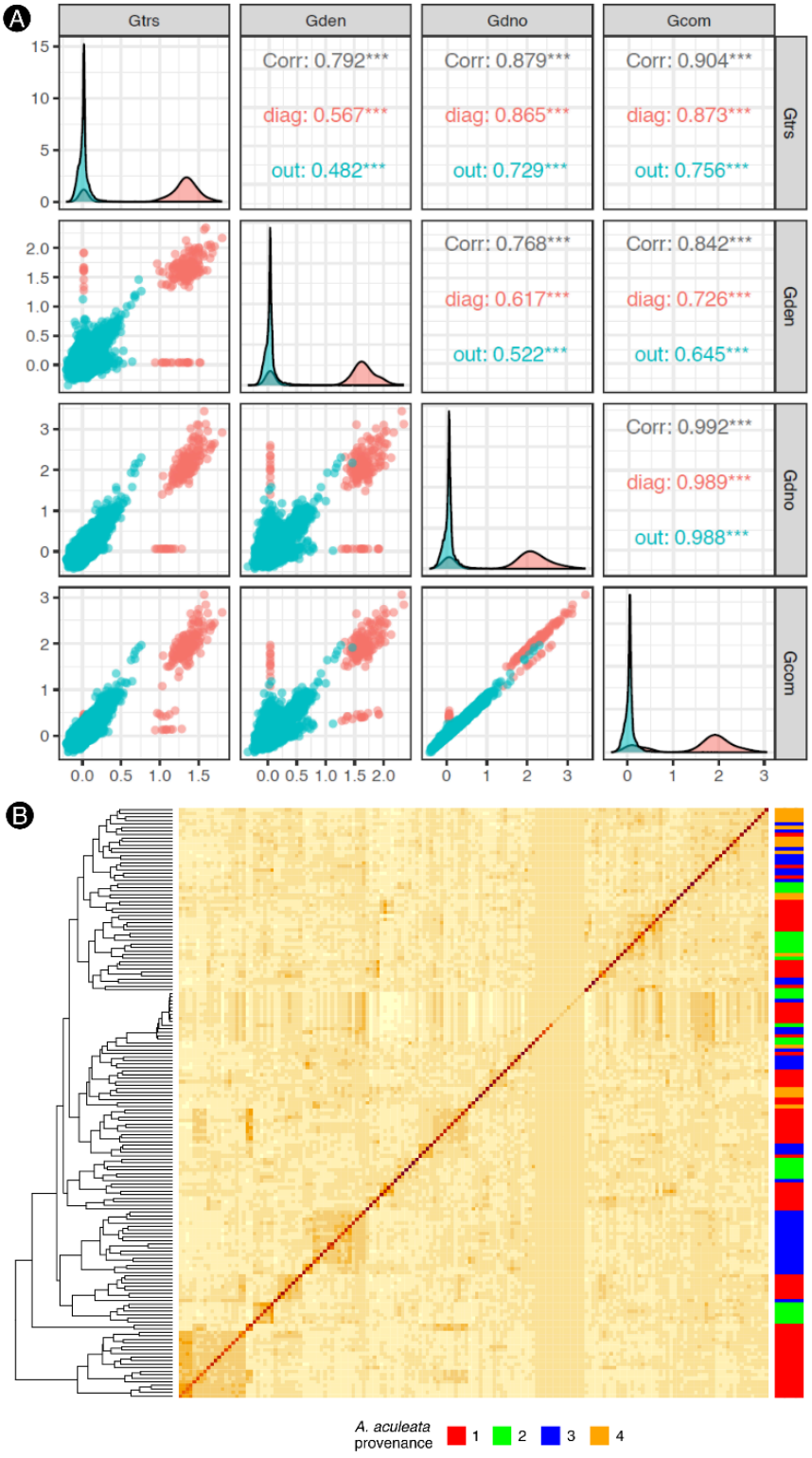

Figure 3

Comparison of genomic relationship matrices and population structure. (A) Pairwise comparisons among the four genomic relationship matrices $Gtrs$ (transcriptome), $Gden$ (oil palm genome), $Gdno$ (de novo pipeline), and $Gcom$ (combined). Correlation coefficients are shown on the diagonals (*** p < 0.001), with within-provenance correlations in red (diag) and between-provenance correlations in blue (out). Scatterplots show relationships among individuals across matrices. (B) Heatmap of the combined genomic matrix ($Gcom$), with hierarchical clustering on the left. Colors on the right indicate provenances (P1 = red, P2 = green, P3 = blue, P4 = orange).

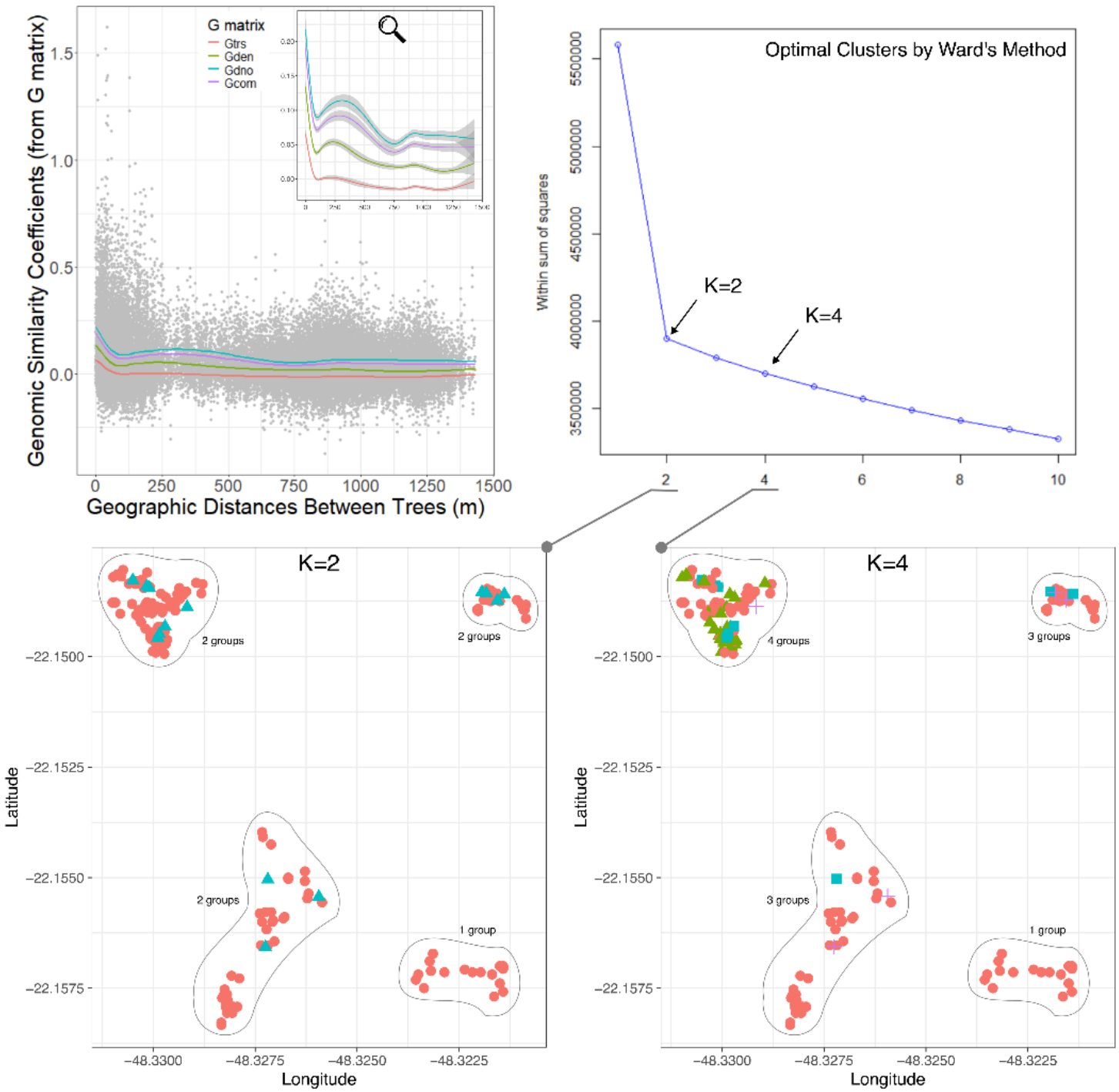

Figure 4

Genomic similarity and spatial clustering of Acrocomia aculeata trees. (Top left) Relationship between genomic

similarity coefficients (from genomic matrices $Gtrs$, $Gden$, $Gdno$ and $Gcom$) and geographic distances between trees (meters). The inset highlights the decay of similarity with distance. (Top right) Determination of the optimal number of clusters using Ward's method applied to the combined genomic matrix ($Gcom$), with K = 2 and K = 4 indicated. (Bottom) Spatial distribution of trees according to cluster assignments (K = 2 and K = 4) based on $Gcom$. Colors indicate clusters, while spatial coordinates correspond to the geographic positions of sampled trees.

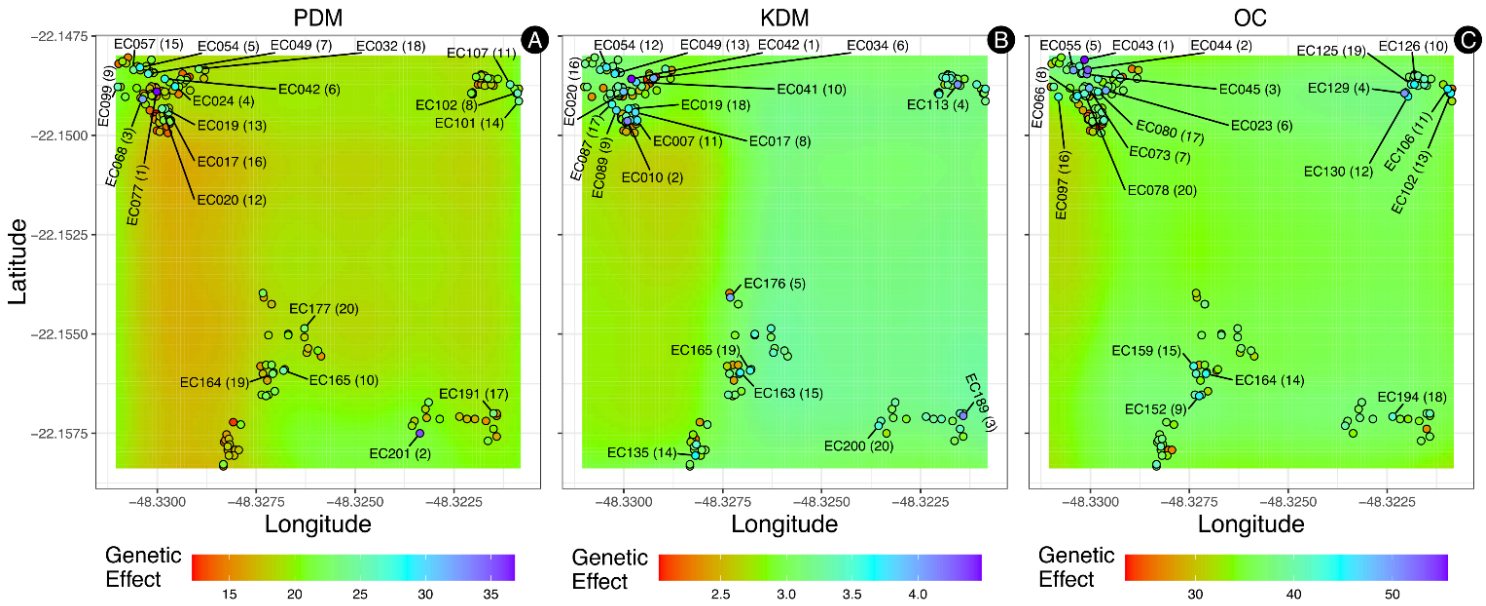

Figure 5

Spatial interpolation of genomic effects for seed collection in Acrocomia aculeata. Genetic effects for pulp dry mass (PDM; A), kernel dry mass (KDM; B), and oil content (OC; C) were obtained using genomic prediction models based on the combined genomic matrix ($Gcom$), and subsequently interpolated with Random Forest. The background gradient represents predicted genetic effect values, with cooler colors indicating higher values. Circles correspond to individual trees, color-coded by their estimated effects. Highlighted labels mark trees with the highest genomic rankings for each trait.