Genet Resour Crop Evol (2024) 71:721–733
https://doi.org/10.1007/s10722-023-01652-2

RESEARCH ARTICLE

Genomic and population characterization of a diversity panel of dwarf and tall coconut accessions from the International Coconut Genebank for Latin America and Caribbean

Allison Vieira da Silva · Emiliano Fernandes Nassau Costa · Leandro Eugenio Cardamone Diniz · Semíramis Rabelo Ramalho Ramos · Roberto Fritsche-Neto

Received: 11 April 2023 / Accepted: 22 June 2023 / Published online: 7 July 2023

Abstract The characterization and construction of genetic diversity panels provide valuable information for developing germplasm conservation strategies and applying advanced breeding techniques. Thus, we performed analysis of diversity and genetic structure in *Cocos nucifera* L. with a collection of dwarf and tall accessions belonging to the International Coconut Genebank for Latin America in the Caribbean. The collection comprises six dwarf accessions (represented by 36 individuals) and six tall (represented by 48 individuals). The analysis of clustering and DAPC performed from a set of 4044 SNP markers showed the existence of three clusters, one of which was formed exclusively by all dwarf coconut accessions. The tall accessions were separated into two distinct clusters, one formed by accessions from regions bathed by the Pacific Ocean (Southeast Asia and Oceania) and another formed by accessions from regions bathed by the Atlantic Ocean (Brazilian coastline and West Africa). The analysis allowed us to observe that the cluster formed by the accessions of dwarf coconut is closer genetically to the cluster formed by the accessions of tall coconut from the Pacific than the others from the Atlantic. Both groups presented similar genetic diversity (GD) values, 0.25 and 0.26, respectively. The inbreeding coefficient (*F*) revealed the presence of greater heterozygosity than expected in the dwarf coconut accessions and the inbreeding in the tall accessions. Consequently, we provide important information for maintaining those accessions in the germplasm bank and for future implementation of genomic-wide association studies (GWAS) and genomic selection (GS) with the evaluated accessions.

Keywords *Cocos nucifera* L. · Genetic variability · Genetic base · Population structure

A. V. da Silva (✉) · R. Fritsche-Neto
Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, Brazil
e-mail: allisonvsagro@usp.br

E. F. N. Costa
Embrapa Tabuleiros Costeiros, Av. Gov. Paulo Barreto de Menezes, 3250, Jardins, Aracaju, Sergipe CEP: 49025-040, Brazil

L. E. C. Diniz
Embrapa Soja, Rodovia Carlos João Strass, S/nº Acesso Orlando Amaral, Distrito de Warta, Londrina, Paraná CEP: 86085-981, Brazil

S. R. R. Ramos
Embrapa Alimentos e Territórios, R. Cincinato Pinto, 348., Maceió, Alagoas CEP: 57020-050, Brazil

R. Fritsche-Neto
Rice Research Station, Louisiana State University Agricultural Center, Baton Rouge, USA

Introduction

The coconut palm (*Cocos nucifera* L.) is a perennial, diploid palm tree (2n = 32) with an estimated genome

size of 2.42 Giga base pairs (Xiao et al. 2017). It is a source of carbohydrates, vitamins, water, oil, and fiber and is used in construction and building, including furniture. Such versatility makes the coconut palm known as the tree of life (Rajesh et al. 2015; Yang et al. 2021). In addition to its various forms of human use, the coconut palm, which occurs along coastal regions of numerous tropical islands, plays a key role in maintaining the ecosystem of these islands (Rajesh et al. 2015).

In 2021, the five largest producers of dried coconut worldwide were Indonesia, the Philippines, India, Sri Lanka, and Brazil (FAOSTAT 2023). It is noted that the main world producers are developing countries, where the crop contributes substantially to the economy, especially the top three in the ranking, which together, in 2021, moved more than 258 million dollars in exports, which corresponds to approximately 64% of all global revenue generated by the coconut exports (FAOSTAT 2023).

The coconut palm is the only species of the genus *Cocos* and is commonly separated into two groups of varieties: the dwarf coconut palm (Nana), self-pollinated, and the tall coconut palm (Typica), which is predominantly cross-pollinated (allogamous). The tall coconut is believed to have emerged on islands in the eastern Pacific and dispersed naturally by sea currents (Clement et al. 2013). On the other hand, the dwarf variety would have evolved from the tall, with great influence of human selection for traits of interest (Perera et al. 2000; Dasanayaka et al. 2009).

The cultivation of the tall coconut is practiced mainly by small farmers. This variety has rapid growth and a long vegetative phase, with the beginning of fruit production between five and seven years after planting and in Brazil, with up to 80 fruits/plant/year (Ribeiro et al. 2012), which are destined mainly for the production of dry coconut and supply the agro-industrial sector. Conversely, the dwarf coconut is the variety most exploited in Brazil to produce coconut water, being more demanding concerning soil and climate conditions, with slow vegetative growth and the beginning of fruit production between two and three years after planting, producing between 150 and 200 fruits per year per plant (Ribeiro et al. 2012), which are intended mainly for the consumption of coconut water *in natura*.

With the marked climate change caused by advancing global warming, scientists are redoubling their attention on the species grown for human food and the stability in growing these species in the face of drastic climate change (Henry 2014). The coconut palm is one of the important species for human nutrition and, consequently, the target of numerous genetic research efforts worldwide. Germplasm banks conserve sources of genes that can define the crop's success in the face of climate change and its consequences. Characterizing the diversity and genetic structure of the accessions conserved in germplasm banks is extremely important for monitoring diversity, allowing breeders enough time and information to develop and apply strategies for maintaining and amplifying the diversity present in the bank and to exploit this diversity in breeding programs.

Some studies have already been conducted to characterize the diversity and genetic structure of coconut accessions conserved in the *International Coconut Genebank for Latin America and the Caribbean* (ICG-LAC) with morphological markers (Sobral et al. 2018, 2019) and microsatellite markers (Loiola et al. 2016). Other work has characterized tall coconut palm's structure and genetic diversity (Ribeiro et al. 2013; Loiola et al. 2016) and dwarf coconut palm (Azevedo et al. 2018; Santos et al. 2020) populations located in Brazil.

Studies of genetic diversity and population structure performed with SNPs markers are becoming increasingly frequent thanks to the robustness of the results generated by the markers (Fischer et al. 2017) when large sets of marks are analyzed, the cheapening and democratization of access to sequencing platforms and advances in bioinformatics (Li et al. 2017). SNPs markers provide relevant information with high reliability regarding the genetic variability of the germplasm bank of a species and how this variability is structured, and are a direct platform for the application of modern breeding techniques. Among these tools, one can mention genomic selection (Bernardo 1994; Meuwissen et al. 2001), which has great potential for reducing the selection cycle, which has a great impact, especially for perennial species, allowing the shortening of generations and enabling the selection of plants still at a reduced size and before fruiting (Kainer et al. 2015; Iwata et al. 2016; Lebedev et al. 2020). For example, GS application promoted a 50% reduction of selection in *Coffea arabica* (Sousa et al. 2019), *Eucalyptus* (Grattapaglia

and Resende 2011), and *Pinus taeda* (Labedev et al. 2020). In addition to reducing the breeding cycle, the use of GS in perennial species promotes savings in physical space and the cost of maintenance in trials since superior genotypes are selected early, and all efforts are focused on the selected individuals (Kainer et al. 2015; Sousa et al. 2019; Fritsche-Neto et al. 2012).

Another tool of great importance is the GWAS, which assists in discovering genes related to adaptive traits of abiotic and biotic stress, directly impacting the repertoire of genetic tools available to breeders in dealing with climate dynamics. Knowledge of population structure is of paramount importance in performing GWAS since the existence of subpopulations in a diverse population can lead to the false association of genetic markers to a phenotype, the association being only a variation in frequency in the marker among some more closely related individuals (Tibbs Cortes et al. 2021). Characterizing the genetic diversity and structure present in germplasm banks of crop species is critical, as it provides the necessary basis for decisions such as selecting parents to be used in crosses that give rise to improved varieties (Park et al. 2021).

Both tools (GS and GWAS) have their application enhanced by the availability of well-characterized panels with broad genetic diversity and the identification and understanding of the population structure and linkage disequilibrium in the accessions. The characterization and construction of genetic diversity panels from SNPs markers have made a great impact in conducting associative genetic studies in major crops such as barley (Rostoks et al. 2006), maize (Yan et al. 2009), wheat (Würschum et al. 2013) and rice (McCouch et al. 2016). Fruit species of great global importance have had germplasm bank accessions characterized, and relevant information has been made available for crop improvement programs such as grape (Emanuelli et al. 2013), peach (Micheletti et al. 2015), apple (Urrestarazu et al. 2016), pear (Li et al. 2019) and mango (Kuhn et al. 2019), for example. Work such as the above, with a large set of SNPs markers, provides valuable information for developing germplasm conservation strategies and applying advanced breeding techniques. In this context, this study aimed to perform the genomic and population characterization of a diversity panel of dwarf and tall coconuts and make its data available to the scientific community in order to boost studies with the species.

Material and methods

Plant material

The study population consisted of six accessions (36 plants) of dwarf and six accessions (48 plants) of tall coconut from the ICG-LAC (International Coconut Genebank for Latin America and the Caribbean), located in Aracaju-SE, Brazil, and maintained by the Brazilian Agricultural Research Corporation (EMBRAPA). The plants selected (Table 1) among the accessions belonging to a working collection originating from the IGC-LAC are promising accessions regarding genetic gains and agronomic traits of interest to farmers.

Genomic characterization

In 2019, DNA was submitted to the University of Wisconsin-Madison Biotechnology Center. DNA concentration was verified using the Quant-iT™ PicoGreen® dsDNA kit (Life Technologies, Grand Island, NY). Libraries were prepared as in Elshire et al. (2011) with minimal modification; in short, 150 ng of

Table 1 List of the dwarf and tall coconut accessions used in the study

Code	Accessions	Origin	N
BYDG	Brazilian yellow dwarf—gramame	Brazil	4
MYD	Malayan yellow dwarf	Malaysia	6
CRD	Cameroon red dwarf	Cameroon	12
BRDG	Brazilian red dwarf—gramame	Brazil	4
MRD	Malayan red dwarf	Malaysia	5
BGDJ	Brazilian green dwarf—jiqui	Brazil	5
BRTPF	Brazilian tall—praia do forte	Brazil	6
WAT	West African tall	Côte d'Ivoire	4
PYT	Polynesian tall	Tahiti	8
RIT	Rennell islands tall	Solomon Islands	12
TONT	Tonga tall	Tonga	9
VTT	Vanuatu tall	Vanuatu	9

The number of plants of each accessory is represented by N

DNA was digested with ApeKI (New England Biolabs, Ipswich, MA), after which barcoded adapters amenable to Illumina sequencing were added by ligation with T4 ligase (New England Biolabs, Ipswich, MA). Finally, 96 adapter-linked samples were pooled and amplified to provide library quantities amenable for sequencing, and adapter dimers were removed by SPRI bead purification. The quality and quantity of the finished libraries were assessed using the Agilent Bioanalyzer High Sensitivity Chip (Agilent Technologies, Inc., Santa Clara, CA) and Qubit® dsDNA HS Assay Kit (Life Technologies, Grand Island, NY), respectively. Libraries were sequenced, targeting 250 million reads on a NovaSeq6000 (Illumina Inc.). Images were analyzed using the standard Illumina Pipeline, version 1.8.2.

Plants were sequenced and genotyped via GBS (Genotyping by Sequencing) by the University of Wisconsin on the Illumina platform. SNP calling was performed based on the reference genome of the dwarf green coconut palm cultivar Catigan (CATD) (NCBI—QRFJ00000000.1). The SNPs markers were filtered using the snpReady package (Granato et al. 2018). Only SNPs with a minimum allele frequency (MAF) of 5% and a call rate of 95% were selected. In the filtering performed for lost data, no individuals with a percentage of lost data greater than 30% were identified. Lost data were imputed using the kinship method. The markers were also filtered for the LD (Linkage Disequilibrium) parameter, the LDs between markers were calculated between 100 kbp intervals from the correlation method, and the 99% threshold was used for filtering the markers.

Statistical-genetic analysis

The genetic structure in the population under study was evaluated based on principal component analysis, genetic distance, and the construction of dendrograms to identify clusters. The analyses were performed considering all individuals as a single population and, in a second scenario, considering two distinct groups, a group formed by the accessions of the dwarf variety and a second group with the tall variety. The genetic diversity was evaluated from parameters such as expected and observed heterozygosity, PIC (polymorphism information content), Nei's genetic diversity, effective population size, and endogamy coefficient. Diversity analyses were generated between brands, genotypes, and populations with the snpReady package (Granato et al. 2018).

Genetic structure analyses were performed using DAPC (Principal Component Discriminant Analysis) (Jombart et al. 2010) and dendrogram construction. The DAPC aims to identify groups based on genetic structuring optimally by reducing the complexity of the data in principal components. The number of groups was not defined before the analysis was performed and was calculated by the K-means method, where several values of k (groups) and several clusters are tested using the Bayesian Information Criterion (BIC). The results are provided in a curve of BIC values as a function of k, where the optimal value of groups is taken as the value at which the curve shows the most pronounced change in behavior. Cluster analyses were performed from phylogenetic trees constructed by the Neighbor-Joining method, without rooting, generated based on Nei (1972) genetic distance with the aid of the ape package (Paradis and Schliep 2019).

Results

Genetic structure

After calling SNPs, a total of 103,057 markers were obtained. With the application of the established parameters for filtering of marks, 4,044 SNPs markers were retained, with which the subsequent analyses were performed. The DAPC analysis was performed without the prior definition of the number of clusters. The number of clusters that best explain the data was obtained from the *find clusters* function by selecting 20 PCs. It was possible to observe that, although the lowest BIC value was obtained with K = 7, from K = 3, the behavior of the line changed dramatically, and the number of clusters was used for the allocation of the accessions (Fig. 1). The analyses performed from the additivity genetic matrix and the dendrograms generated also served as a basis for identifying the number of clusters that best explain the distribution of the accessions.

In the scatterplot generated from the DAPC analysis, the three clusters identified are distributed in an isolated manner, without overlap between clusters, reinforcing the genetic divergence between them (Fig. 2).

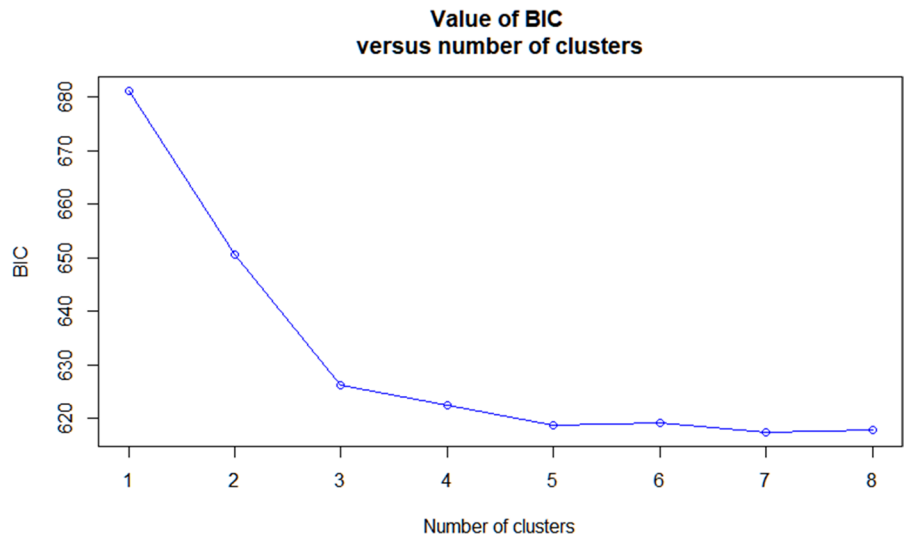

Fig. 1 Number of clusters (K) identified based on BIC. The accessions were optimally clustered with the value of K = 3

Fig. 2 Scatter plot of the 12 accessions (84 plants) of dwarf and tall coconut, generated by Principal Component Discriminant Analysis (DAPC), from the set of 4044 SNPs. The DAPC analysis retained 35 principal components (PCs) and two discriminant functions to evaluate the relationship between clusters

The distribution of the coconut accessions in the DAPC analysis was similar to the distribution in the dendrograms constructed from the Euclidean distance and the Jaccard coefficient. All dwarf coconut accessions formed the first cluster (in bold). The tall coconut accessions, in turn, were separated into two clusters. A second cluster (in italic) was formed by accessions from the WAT and BRTPF populations. Only one of the accessions from the TONT population was allocated to the second cluster. All other populations and tall coconut accessions were allocated to a third cluster, represented by bolditalic (Table 2).

Based on the DAPC analysis, a bar graph was generated to identify the probability of grouping the accessions to the three identified clusters (Fig. 3). The analysis allowed the observation of well-defined groups without mixtures. None of the accessions showed the probability of grouping with any other cluster than the one in which the DAPC analysis allocated them.

Table 2 Distribution of the accessions in the identified clusters according to group and representative cluster coloration

Group	Accessions					
Dwarf	**BYDG**	**MYD**	**CRD**	**BRDG**	**MRD**	**BGDJ**
Tall	*BRTPF*	*WAT*	***PYT***	***RIT***	***TONT***	***VTT***

The first cluster is identified in bold, the second in italic, and the third in bolditalic

BYDG Brazilian Yellow Dwarf—Gramame, *MAYD* Malayan Yellow Dwarf, *CRD* Cameroon Red Dwarf, *BRDG* Brazilian Red Dwarf—Gramame, *MRD* Malayan Red Dwarf, *BGDJ Brazilian* Green Dwarf—Jiqui, *BRTPF* Brazilian Tall—Praia do Forte, *WAT* West African Tall, *PYT* Polynesian Tall, *RIT* Rennell Islands Tall, *TONT* Tonga Tall, *VTT* Vanuatu Tall

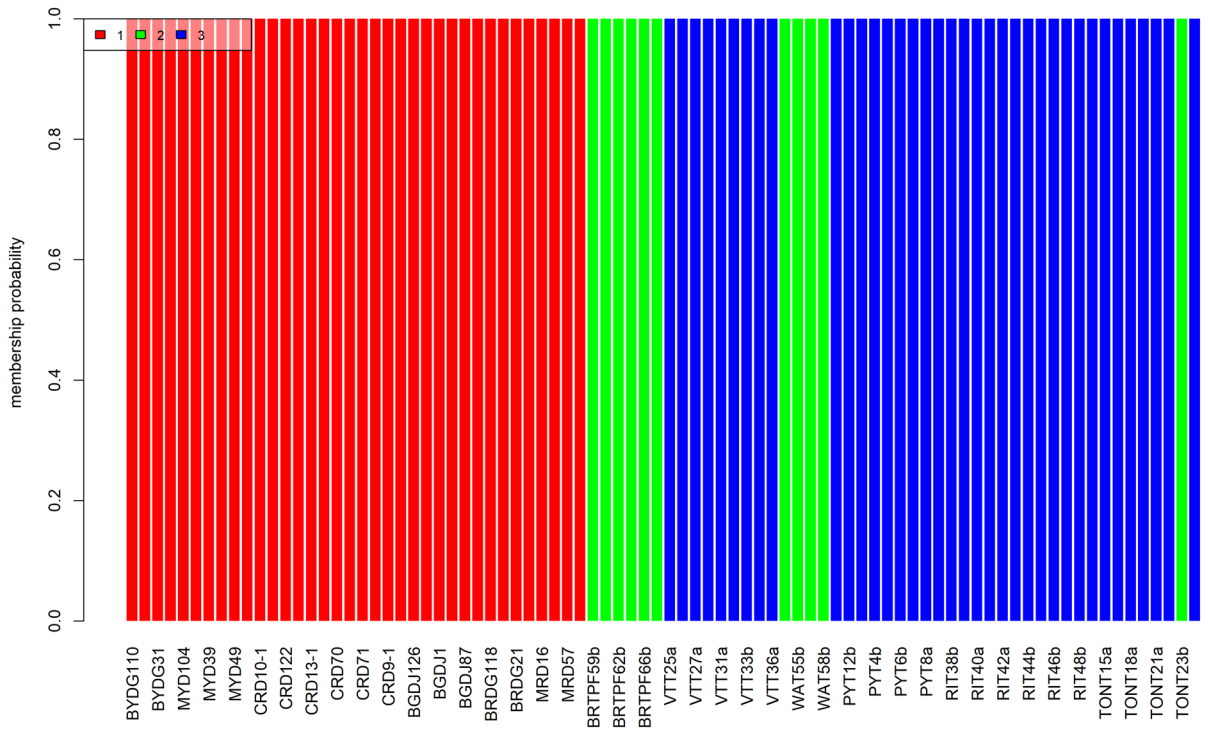

Fig. 3 Bar graph with the probability of grouping for each of the 84 individuals among the three identified clusters (K = 3)

Three large clusters were identified in the dendrogram generated from the Euclidean distance (Fig. 4). One group was formed by tall coconut plants of the WAT and BRTPF accessions, a second group with all the dwarf coconut accessions, and a third group with the other tall coconut accessions.

In the dendrogram constructed from Jaccard's distance, the three-way structuring was maintained in conjunction with the distribution of the evaluated plants, although minor changes appeared in the dendrogram architecture (Fig. 5).

Genetic diversity

As observed in Table 3, among the three scenarios in which the genetic diversity analyses were performed, 12 accessions (84 plants) showed higher genetic diversity (GD) than the two groups analyzed separately, with a value of 0.31. Although the difference is not significant between the two groups, the tall coconut accessions had a higher value for genetic diversity (0.26) compared to the

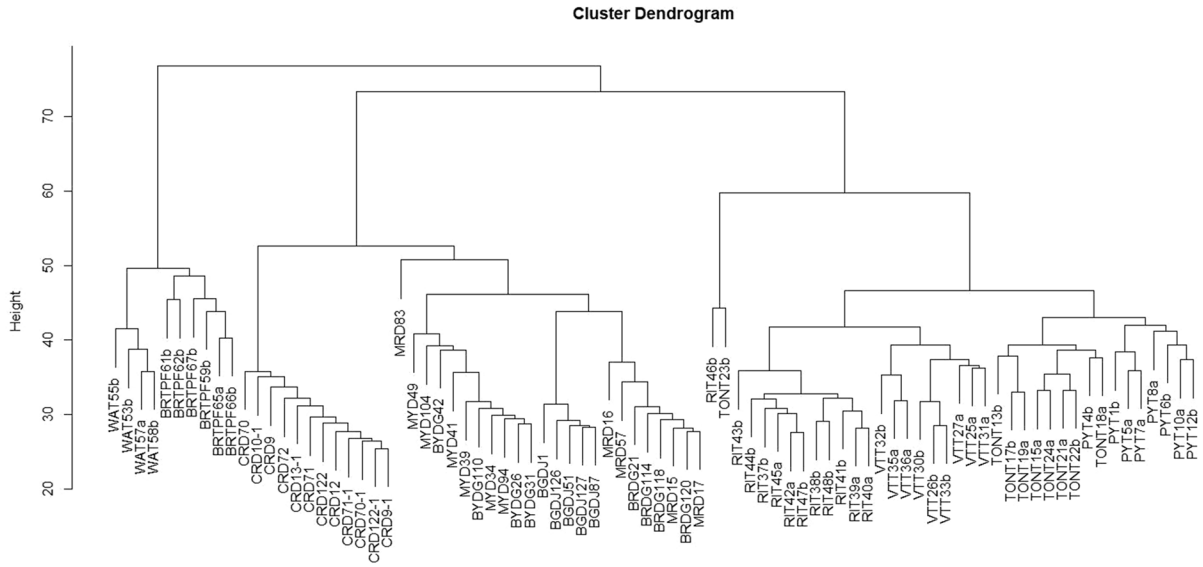

Fig. 4 Dendrogram constructed from the Euclidean distance for cluster analysis of the 12 coconut accessions (84 plants)

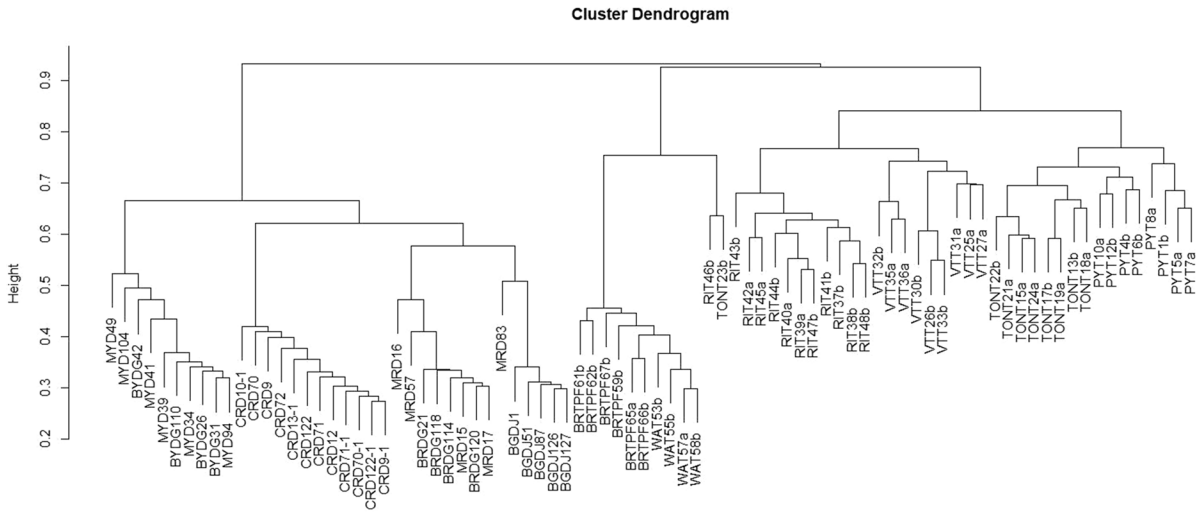

Fig. 5 Dendrogram generated from Jaccard's similarity index to evaluate the clustering among the 12 accessions (84 plants) of coconut evaluated

group formed by dwarf coconut accessions (0.25). The values for PIC showed similar behavior to the values for GD, with the general group showing the highest value for the variable (0.25), followed by the group formed by tall coconut accessions (0.22), and the group formed by dwarf coconut trees showed the lowest value (0.20).

Linkage disequilibrium

The linkage disequilibrium (LD) analysis performed with the dwarf coconut accessions demonstrated the presence of large linkage blocks. Although some of the chromosomes showed low coverage by marks, it was possible to observe that the general pattern for linkage disequilibrium is repeated among

Table 3 Mean values of genetic diversity parameters evaluated with all accessions, forming only one group (general), with the dwarf coconut accessions and the tall coconut accessions separately

GD	PIC	MAF	Ho	F	Ne
General					
0.31	0.25	0.21	0.22	0.27	152.97
Dwarf					
0.25	0.20	0.20	0.29	−0.16	−112.7
Tall					
0.26	0.22	0.18	0.17	0.35	68.2

GD Ney genetic diversity, *PIC* polymorphism content, *MAF* minor allele frequency, *Ho* observed heterozygosity, *F* inbreeding coefficient, *Ne* effective population size

chromosomes, showing attenuated initial drop or almost no drop along the chromosome (Fig. 6).

When observing the linkage disequilibrium analysis from the tall coconut individuals (Fig. 7), it is possible to observe that these had greater coverage regarding the distribution of the marks along the chromosomes, except for chromosomes 15 and 17, which presented less coverage. In chromosomes with more dense coverage, verifying the presence of linkage blocks becomes confusing, as in chromosomes two, three, and six. We can see that the marks have no defined distribution pattern in chromosomes with less dense coverage, as in chromosomes one, five, and nine. The drop in linkage disequilibrium, represented by the red line, showed the same pattern on most chromosomes.

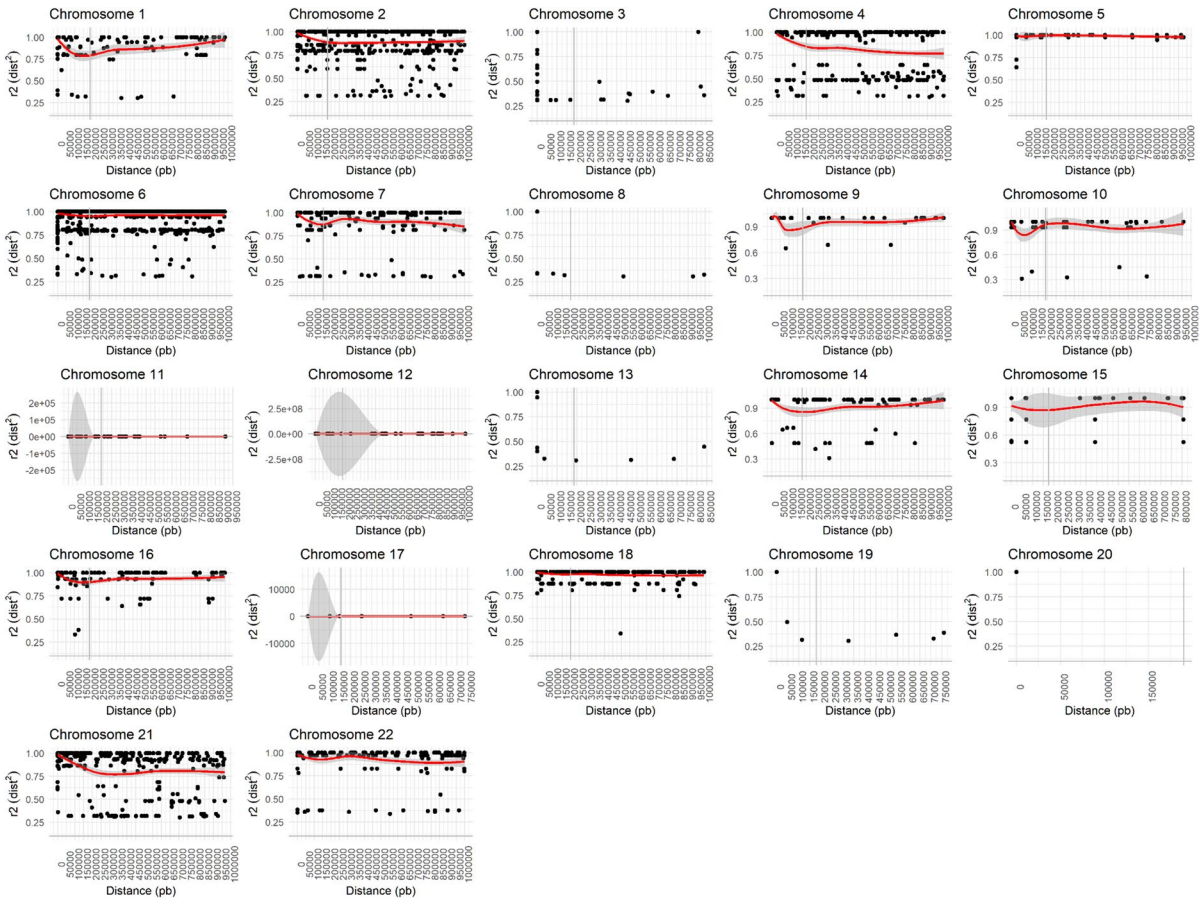

Fig. 6 Graph of linkage disequilibrium (r^2) on the vertical axis, and distance between SNPs markers, on the horizontal axis, along each of the chromosomes sampled in the 36 dwarf coconut accessions

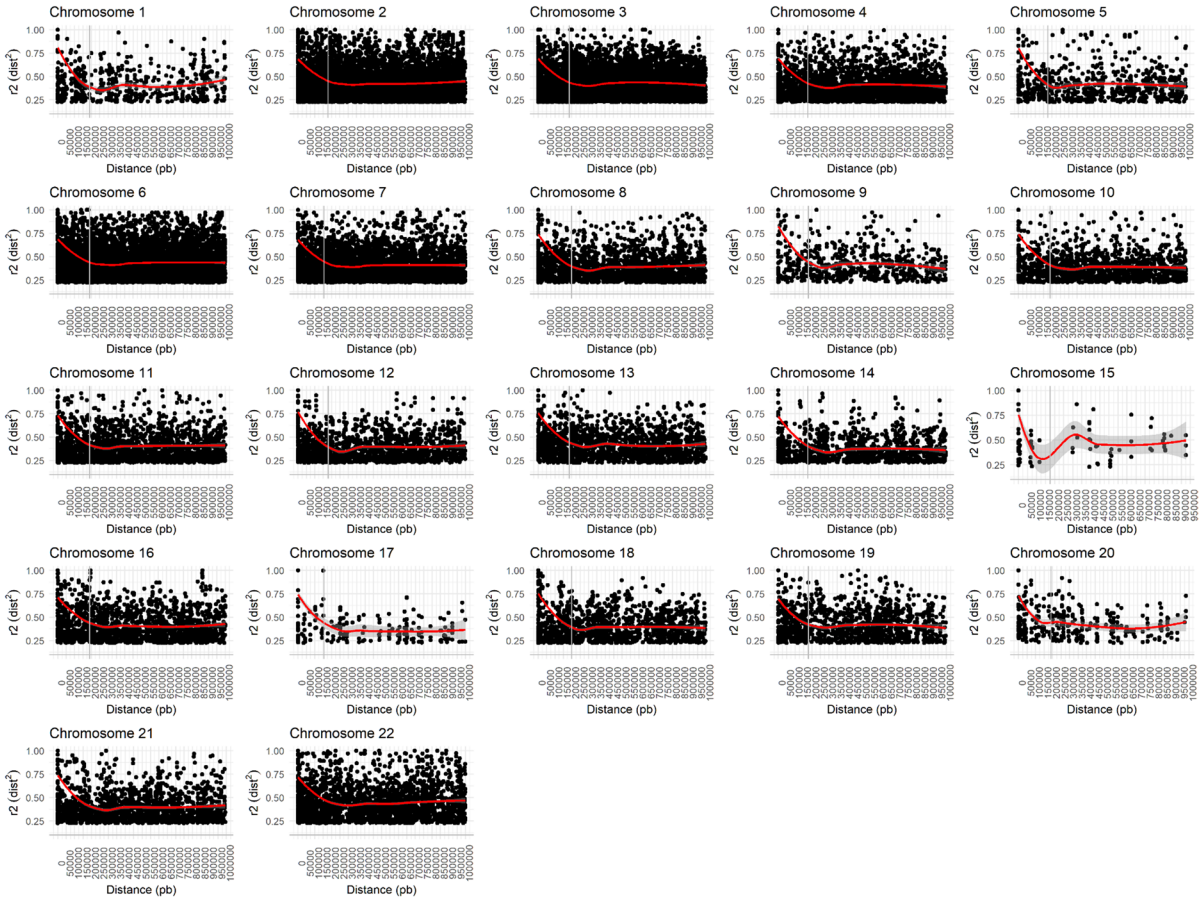

Fig. 7 Graph of linkage disequilibrium (r^2) on the vertical axis and distance between SNPs markers, on the horizontal axis, along each of the chromosomes sampled in the 48 tall coconut accessions

Discussion

The structuring observed in tall coconut individuals is in agreement with the pattern observed in other works (Perera et al. 2003; Gunn et al. 2011; Loiola et al. 2016; Muñoz-Pérez et al. 2022) that also identified structuring into different groups according to regions bordering the Atlantic Ocean and regions bordering the Pacific Ocean. This evidence suggests that the domestication of coconut occurred independently in the Atlantic and Pacific (Perera et al. 2003; Gunn et al. 2011). Furthermore, the genetic proximity observed between individuals from Brazilian and West African populations reinforces the theory of the introduction of the coconut palm to the Atlantic Coast of the Americas by Portuguese expeditions from the Cape Verde islands (Clement et al. 2013; Loiola et al. 2016).

All individuals of dwarf coconut were allocated to a single group, showing that the individuals have a narrow genetic base and are strongly related. Santos et al. (2020), in a study conducted with populations of dwarf coconut collected in producing areas of Brazil and analysis based on SNP markers, observed the low differentiation between populations, in agreement with our results. Gunn et al. (2011) evaluated the diversity and genetic structure of dwarf and tall coconut individuals throughout the species' geographical dispersion and verified the low genetic differentiation among the dwarf coconut accessions evaluated. Although they formed a distinct group, the dwarf coconut individuals are genetically closer to the tall

coconut individuals of Pacific origin. This genetic proximity has also been identified in other works using molecular markers and reinforces the theory that the dwarf coconut palm was domesticated in Southeast Asia (Perera et al. 2003; Gunn et al. 2011; Muñoz-Perez et al. 2022). In this context, studies of genetic diversity and structure conducted with the dwarf and tall coconut varieties together suggest that the dwarf coconut variety originated from artificial selection performed on a small group of tall coconuts and evidence of the presence of greater genetic diversity within the tall coconut variety (Perera et al. 2000, 2003).

The mean value of genetic diversity of the dwarf coconut individuals evaluated (0.25) is similar to the results obtained by Jean Nöl et al. (2011) and by Gunn et al. (2011) when they evaluated populations of dwarf coconut using SSR markers, obtaining a mean value of genetic diversity around 0.218 and 0.270, respectively. However, the tall coconut individuals showed a mean value for genetic diversity of 0.26. Because they are autogamous plants, the low value of genetic diversity is within the expected for the dwarf coconut individuals, the opposite happens with the tall coconut individuals, where a higher mean value of genetic diversity would be expected because they are allogamous plants (Perera et al. 2003; Ribeiro et al. 2013). Loiola et al. (2016) evaluated 90 individuals of tall coconut corresponding to nine accessions also from the ICG-LAC, some of them present in this work (BRTPF, WAT, RIT, TONT, and VTT), through SSR markers and observed higher mean values of genetic diversity (0.47). The divergence found in the diversity values of the tall coconut palm may come from the artificial selection pressure for the composition of the working collection evaluated in our work. Loiola et al. (2016) also identified significant values for the endogamy coefficient for two of the accessions evaluated (0.32 and 0.34), one of them also from WAT, which is similar to the average value found in this work for the tall coconut individuals (0.35).

Identifying the LD structure contributes to the rapid identification and selection of alleles of agronomic interest in obtaining improved varieties in breeding programs (Yan et al. 2009). LD is one of the diversity parameters that have a direct influence on the application of GWAS and in making accurate predictions because of its ability to add bias to association analyses (Robbins et al. 2011; Porto-Neto et al. 2014). Autogamous plants generally show a more attenuated LD decay pattern due to the predominance of loci in homozygosity and the presence of large linkage blocks throughout the genome, reducing the frequency of recombination events (Vos et al. 2017). We could observe this pattern in the LD analysis performed with the dwarf coconut individuals, predominantly autogamous plants. Gene recombination events typically occur more frequently in allogamous species, which is reflected in the rapid drop in the linkage disequilibrium curve in these species (Vos et al. 2017). Significant estimates of linkage disequilibrium in a given population indicate the presence of evolutionary pressure on the population, such as inbreeding, gene flow, genetic drift, mutation, and natural selection (Zhu et al. 2015). We observed that tall coconut individuals showed a pattern of LD decay that can be considered a faint decline, considering that tall coconut individuals are allogamous plants. It may be related, as well as the low values presented for the GD index and high values of the inbreeding coefficient, to genetic drift mechanisms acting on these individuals. Finally, significant estimates of linkage disequilibrium in a given population indicate the presence of evolutionary pressure on that population, such as inbreeding, gene flow, genetic drift, mutation, and natural selection (Zhu et al. 2015).

For genomic prediction, the models are developed from a set of genotyped and phenotyped individuals, which form a training population (TP), and applied to the breeding population, with the presence of individuals or offspring from the training population, to obtain the estimated breeding values of the individuals in the breeding population (Desta and Ortiz 2014; Crossa et al. 2017; Xu et al. 2020). With the model, it is possible to rank individuals for a trait of interest without the need for knowledge of the phenotype of these individuals, using only genotypic data and Mendelian sampling of offspring to perform the estimates of breeding values in the target population (Kwong et al. 2017). Selection based only on genotypic data can be performed in the early stages of development, accelerating the breeding program and promoting the increment of annual gain from the shortening of the crop cycle (Xu et al. 2020).

The existence of genetic diversity for selection and the composition of the training population and validation population has a direct influence on the accuracy of genomic prediction, using a genetically divergent

training population concerning the validation population can cause overestimated accuracy values (Wray et al. 2013). Therefore, structuring between populations is important in genomic selection studies and directly influences the composition of the training population (Guo et al. 2014; Isidro et al. 2015; Olatoye et al. 2020).

Our work evidences the presence of genetic structure among the tall coconut individuals evaluated and the genetic differentiation among the dwarf and tall coconut individuals. We also provide relevant information on the parameters of genetic diversity that will assist in curating the germplasm bank and in the maintenance and expansion of this diversity from targeted crosses based on the genetic structure observed. We make available here new information of extreme importance and usefulness in conducting future work aimed at the application of advanced quantitative genetics tools such as GS and GWAS.

Acknowledgements "The author(s) utilized the University of Wisconsin—Madison Biotechnology Center's DNA Sequencing Facility (Research Resource Identifier—RRID:SCR_017759) to generate GBS libraries and sequence GBS libraries. The UWBC is a Licensed Service Provider for internal and external clients, providing GBS services under license from Keygene N.V. which owns patents and patent applications protecting its Sequence Based Genotyping technologies."

Author contribution All author contributed with the study development. Marial preparation and data collection were performed by EFNC, LECD and SRRR. Data analysis were performed by RF-N and AVS. The manuscript was written by all the authors since the first draft till the final version. All authors read and approved the final manuscript.

Funding This work was supported by EMBRAPA (Project number: 03.15.00.127.00.00 and 01.06.01.007.06.00).

Data availability "The raw SNPs markers generated during and analyzed during the current study are available in the Mendeley Data repository, https://data.mendeley.com/datasets/42p5djgtjs."

Declarations

Conflict of interest The authors have non-financial interests to disclose.

References

Azevedo AON, de Azevedo CDO, Santos PHAD et al (2018) Selection of legitimate dwarf coconut hybrid seedlings using DNA fingerprinting. Crop Breed Appl Biotechnol 18:409–416. https://doi.org/10.1590/1984-70332018v18n4a60

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25. https://doi.org/10.2135/cropsci1994.0011183X003400010003x

Clement CR, Zizumbo-Villarreal D, Brown CH et al (2013) Cocoteros en las Américas. Bot Rev 79:342–370. https://doi.org/10.1007/s12229-013-9121-z

Crossa J, Pérez-Rodríguez P, Cuevas J et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Dasanayaka PN, Everard JMDT, Karunanayaka EH, Nandadasa HG (2009) Analysis of coconut (*Cocos nucifera* L.) diversity using microsatellite markers with emphasis on management and utilisation of genetic resources. J Natl Sci Found Sri Lanka 37:99–109. https://doi.org/10.4038/jnsfsr.v37i2.1065

Desta ZA, Ortiz R (2014) Genomic selection: Genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601. https://doi.org/10.1016/j.tplants.2014.05.006

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

Emanuelli F, Lorenzi S, Grzeskowiak L et al (2013) Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol 13:1–17. https://doi.org/10.1186/1471-2229-13-39

Fischer MC, Rellstab C, Leuzinger M et al (2017) Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in Arabidopsis halleri. BMC Genomics 18:1–15. https://doi.org/10.1186/s12864-016-3459-7

Food and Agriculture Organization of the United Nations (2023) FAOSTAT statistical database. https://www.fao.org/faostat/en/#data/TCL. Accessed 18 Jan 2023

Fritsche-Neto R, Resende MDV, Miranda GV, DoVale JC (2012) Seleção genômica ampla e novos métodos de melhoramento do milho. Rev Ceres 59:794–802. https://doi.org/10.1590/s0034-737x2012000600009

Granato ISC, Galli G, de Oliveira Couto EG et al (2018) snpReady: a tool to assist breeders in genomic analysis. Mol Breed. https://doi.org/10.1007/s11032-018-0844-8

Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255. https://doi.org/10.1007/s11295-010-0328-4

Gunn BF, Baudouin L, Olsen KM (2011) Independent origins of cultivated coconut (*Cocos nucifera* L.) in the old world tropics. PLoS ONE. https://doi.org/10.1371/journal.pone.0021143

Guo Z, Tucker DM, Basten CJ et al (2014) The impact of population structure on genomic prediction in stratified

populations. Theor Appl Genet 127:749–762. https://doi.org/10.1007/s00122-013-2255-x
Henry RJ (2014) Genomics strategies for germplasm characterization and the development of climate resilient crops. Front Plant Sci 5:1–4. https://doi.org/10.3389/fpls.2014.00068
Isidro J, Jannink JL, Akdemir D et al (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158. https://doi.org/10.1007/s00122-014-2418-4
Iwata H, Minamikawa MF, Kajiya-Kanegae H et al (2016) Genomics-assisted breeding in fruit trees. Breed Sci 66:100–115. https://doi.org/10.1270/jsbbs.66.100
Jean Nöl KK, Edmond KK, Konan KJL, Eugene KK (2011) Microsatellite gene diversity within Philippines dwarf coconut palm (*Cocos nucifera* L.) resources at Port-Bouët. Côte D'ivoire Sci Res Essays 6:5986–5992. https://doi.org/10.5897/SRE11.877
Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94. https://doi.org/10.1186/1471-2156-11-94
Kainer D, Lanfear R, Foley WJ, Külheim C (2015) Genomic approaches to selection in outcrossing perennials: focus on essential oil crops. Theor Appl Genet 128:2351–2365. https://doi.org/10.1007/s00122-015-2591-0
Kuhn DN, Dillon N, Bally I et al (2019) Estimation of genetic diversity and relatedness in a mango germplasm collection using SNP markers and a simplified visual analysis method. Sci Hortic (amsterdam) 252:156–168. https://doi.org/10.1016/j.scienta.2019.03.037
Kwong QB, Ong AL, Teh CK et al (2017) Genomic selection in commercial perennial crops: applicability and improvement in oil palm (Elaeis guineensis Jacq.). Sci Rep 7:1–9. https://doi.org/10.1038/s41598-017-02602-6
Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic selection for forest tree improvement: methods, achievements and perspectives. Forests 11:1–36. https://doi.org/10.3390/f11111190
Li PE, Lo CC, Anderson JJ et al (2017) Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res 45:67–80. https://doi.org/10.1093/nar/gkw1027
Li X, Singh J, Qin M et al (2019) Development of an integrated 200K SNP genotyping array and application for genetic mapping, genome assembly improvement and genome wide association studies in pear (Pyrus). Plant Biotechnol J 17:1582–1594. https://doi.org/10.1111/pbi.13085
Loiola CM, Azevedo AON, Diniz LEC et al (2016) Genetic relationships among tall coconut palm (*Cocos nucifera* L.) accessions of the international coconut Genebank for Latin America and the Caribbean (ICG-LAC), evaluated using microsatellite markers (SSRs). PLoS ONE 11:1–11. https://doi.org/10.1371/journal.pone.0151309
McCouch SR, Wright MH, Tung CW et al (2016) Open access resources for genome-wide association mapping in rice. Nat Commun. https://doi.org/10.1038/ncomms10532
Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819
Micheletti D, Dettori MT, Micali S et al (2015) Whole-genome analysis of diversity and SNP-major gene association in peach germplasm. PLoS ONE 10:1–19. https://doi.org/10.1371/journal.pone.0136803
Muñoz-Pérez JM, Cañas GP, López L, Arias T (2022) Genome-wide diversity analysis to infer population structure and linkage disequilibrium among Colombian coconut germplasm. Sci Rep 12:1–11. https://doi.org/10.1038/s41598-022-07013-w
Nei M (1972) Genetic distance between populations. Am Nat 106:283–292
Olatoye MO, Clark LV, Labonte NR et al (2020) Training population optimization for genomic selection in miscanthus. G3 genes. Genomes Genet 10:2465–2476. https://doi.org/10.1534/g3.120.401402
Paradis E, Schliep K (2019) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528. https://doi.org/10.1093/bioinformatics/bty633
Park S, Kumar P, Shi A, Mou B (2021) Population genetics and genome-wide association studies provide insights into the influence of selective breeding on genetic variation in lettuce. Plant Genome 14:1–12. https://doi.org/10.1002/tpg2.20086
Perera L, Russell JR, Provan J, Powell W (2000) Use of microsatellite DNA markers to investigate the level of genetic diversity and population genetic structure of coconut (*Cocos nucifera* L.). Genome 43:15–21. https://doi.org/10.1139/gen-43-1-15
Perera L, Russell JR, Provan J, Powell W (2003) Studying genetic relationships among coconut varieties/populations using microsatellite markers. Euphytica 132:121–128. https://doi.org/10.1023/A:1024696303261
Porto-Neto LR, Kijas JW, Reverter A (2014) The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genet Sel Evol 46:1–5. https://doi.org/10.1186/1297-9686-46-22
Rajesh MK, Sabana AA, Rachana KE et al (2015) Genetic relationship and diversity among coconut (*Cocos nucifera* L.) accessions revealed through SCoT analysis. 3 Biotech 5:999–1006. https://doi.org/10.1007/s13205-015-0304-7
Ribeiro FE, Baudouin L, Lebrun P et al (2013) Genetic diversity in Brazilian tall coconut populations by microsatellite markers. Crop Breed Appl Biotechnol 13:356–362. https://doi.org/10.1590/s1984-70332013000400006
Ribeiro FE, Costa EFN, Aragão W (2012) Árvore do conhecimento: coco. https://www.agencia.cnptia.embrapa.br/gestor/coco/Abertura.html. Accessed 13 Aug 2021
Robbins MD, Sim SC, Yang W et al (2011) Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. J Exp Bot 62:1831–1845. https://doi.org/10.1093/jxb/erq367
Rostoks N, Ramsay L, MacKenzie K et al (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varietes. Proc Natl Acad Sci USA 103:18656–18661. https://doi.org/10.1073/pnas.0606133103
Santos PHAD, Venâncio TM, dos Santos PHD et al (2020) Genotyping-by-sequencing technology reveals directions for coconut (*Cocos nucifera* L.) breeding strategies

for water production. Euphytica. https://doi.org/10.1007/s10681-020-02582-1
Sobral KMB, De Queiroz MA, Da Silva Ledo CA et al (2018) Genetic diversity assessment among tall coconut palm. Rev Caatinga 31:28–39. https://doi.org/10.1590/1983-21252018v31n104rc
Sobral KMB, De Queiroz MA, Neto IDSL et al (2019) Is there genetic variability in dwarf coconut accessions preserved in brazil? Rev Caatinga 32:52–61. https://doi.org/10.1590/1983-21252019v32n106rc
Sousa TV, Caixeta ET, Alkimim ER et al (2019) Early selection enabled by the implementation of genomic selection in coffea arabica breeding. Front Plant Sci 9:1–12. https://doi.org/10.3389/fpls.2018.01934
Tibbs Cortes L, Zhang Z, Yu J (2021) Status and prospects of genome-wide association studies in plants. Plant Genome 14:1–17. https://doi.org/10.1002/tpg2.20077
Urrestarazu J, Denancé C, Ravon E et al (2016) Analysis of the genetic diversity and structure across a wide range of germplasm reveals prominent gene flow in apple at the European level. BMC Plant Biol 16:1–20. https://doi.org/10.1186/s12870-016-0818-0
Vos PG, Paulo MJ, Voorrips RE et al (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theor Appl Genet 130:123–135. https://doi.org/10.1007/s00122-016-2798-8
Wray NR, Yang J, Hayes BJ et al (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14:507–515. https://doi.org/10.1038/nrg3457
Würschum T, Langer SM, Longin CFH et al (2013) Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. Theor Appl Genet 126:1477–1486. https://doi.org/10.1007/s00122-013-2065-1
Xiao Y, Xu P, Fan H et al (2017) The genome draft of coconut (*Cocos nucifera* L.). Gigascience 6:1–11. https://doi.org/10.1093/gigascience/gix095
Xu Y, Liu X, Fu J et al (2020) Enhancing Genetic gain through genomic selection: from livestock to plants. Plant Commun 1:100005. https://doi.org/10.1016/j.xplc.2019.100005
Yan J, Shah T, Warburton ML et al (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE. https://doi.org/10.1371/journal.pone.0008451
Yang Y, Bocs S, Fan H et al (2021) Coconut genome assembly enables evolutionary analysis of palms and highlights signaling pathways involved in salt tolerance. Commun Biol. https://doi.org/10.1038/s42003-020-01593-x
Zhu X, Dong L, Jiang L et al (2015) Constructing a linkage-linkage disequilibrium map using dominant-segregating markers. DNA Res 23:1–10. https://doi.org/10.1093/dnares/dsv031