First chromosome scale genome of *Acrocomia aculeata*

Matheus Scaketti[1]; Caroline Bertocco Garcia[2]; João Victor da Silva Rabelo Araujo[1]; Igor Araujo Santos de Carvalho[2]; Jonathan Morales Marroquin[2;] Diego Mauricio Riaño Pachón[2]; Durval Dourado Neto[2]; Carlos Alberto Labate[2]; Renato Vicentini dos Santos[1]; José Baldin Pinheiro[2]; Suelen Alves Vianna[4]; Carlos Colombo[5]; Maria Imaculada Zucchi[126]

1. Biology Institute, State University of Campinas - UNICAMP, Campinas, São Paulo, Brazil.
2. Genetics Department, University of São Paulo - ESALQ/USP, Piracicaba, São Paulo, Brazil.
3. Center for Nuclear Energy in Agriculture - CENA, University of São Paulo - ESALQ/USP, Piracicaba, São Paulo, Brazil.
4. Acelen Renovaveis, São Paulo, São Paulo, Brazil
5. Agronomic Institute of Campinas, Campinas, São Paulo, Brazil
6. Secretariat of Agriculture and Food Supply of São Paulo State, APTA, Piracicaba, São Paulo, Brazil

Abstract

The Arecaceae family comprises economically, and ecologically significant palm species widely distributed across tropical and subtropical regions. Among them, *Acrocomia aculeata*, commonly known as Macauba, has gained attention due to its high oil yield, environmental adaptability, and potential applications in sustainable agriculture and bioenergy production. This study presents the first chromosome scale genome assembly of *A. aculeata* and the first reference genome of the genus, using Oxford Nanopore Technologies sequencing, PacBio HiFi and Hi-C proximity ligation data. The genome covers 1.94 Gbp in the 15 pseudochromosomes, with N50 of 143.43 Mbp, achieving a base quality QV of 76.4 and a k-mer completeness of 90.58% of PacBio HiFi data. Interspersed repetitive regions represents 77.36% of the total genome, with long terminal repeat (LTRs) accounting for 54.98%. A total of 28.204 protein-coding genes were predicted. The genome was validated with BUSCO analysis, achieving a high completeness of 98.3% for predicted proteins. The first chromosome scale genome of Macauba can provide new insights and breakthroughs for studies of conservation genomics and plant breeding programs.

Background & Summary

Acrocomia aculeata, commonly known as Macauba, macaw palm, or coyol, is a palm species in the order Arecales order within angiosperm. Native from Mexico to northern Argentina and the Caribbean, *A. aculeata* is highly adaptable to various environments[1]. In

Brazil, it thrives in the Cerrado, Atlantic Forest, and Pantanal phytogeographic domains, particularly in pasturelands and degraded areas, where it plays a role in ecological restoration[2]. This adaptability position it as a strategic species for sustainable agriculture and reforestation programs throughout Latin America[3].

One of the most valuable traits of Macauba is its high oil yield, comparable to *Elaeis guineensis*. The mesocarp oil content can range from 20% to 75% and is rich in monounsaturated fatty acids, predominantly oleic acid (~49.3%), while the kernel oil ranges from 50% to 60% and contains high levels of lauric acid (~45.1%)[4]. This composition allows for diverse applications in biodiesel production, food products, and cosmetics, strengthening Macauba's role in sustainable economic development[5]. As a consequence, it has received attention as an alternative and sustainable source of biofuels and high value bioproducts, presenting potential benefits for reducing environmental impact and promoting renewable energy[6].

In this context, breeding initiatives for Macauba are gaining strength due to the species' potential. Until now, there are some studies[7–10] showing that Macauba can be used in a selective breeding strategy. Although all these efforts, genomics resources for palm species remain scarce, with only 21 published genomes within Arecales, for 9 different species[11], with just *Elaeis oleifera* being native to Brazil[1].

To help fill this gap, we report the first chromosome scale of one individual of the important socio-economical species *A. aculeata* using Oxford Nanopore Technologies (ONT), Pacific Biosciences HiFi (PacBio) and Hi-C technology. This is also the first genome of the genus and can provide new insights for the studies of the genetic structure and the evolution of the species, allowing new breakthroughs for both conservation genomics and plant breeding programs.

Methods

Sample collection, library construction and sequencing

A young sapling of *A. aculeata was* collected from IAC (Agronomic Institute of Campinas). Nearly 9,5 g of young leaves (30 leaves) of an approximated 24-month-old palm tree were freshly collected after a 48-hour dark treatment, immediately frozen in liquid nitrogen, and stored at –80 °C until further processing.

High-molecular-weight (HMW) DNA was extracted using a protocol based on Jones et al.[12] and the yielded HMW DNA was used for all ONT libraries. HMW DNA samples were purified with AMPure XP beads (Beckman Coulter Inc., CA, United States), and DNA fragments ranging from 10 to 25 Kbp were selectively depleted using the SRE kits from PacBio and ONT. Quality control was performed using a NanoDrop® ND-1000 spectrophotometer (Thermo Fisher Scientific, MA, USA), and the A260/280 and A260/230 ratios were assessed to determine DNA purity. Genomic DNA concentration was confirmed using a Qubit® 3.0 fluorometer (Invitrogen, CA, USA) with the dsDNA BR Assay Kit (Thermo Fisher Scientific, MA, USA, Q32850). The integrity and size of high-molecular-weight (HMW) DNA fragments were evaluated using the 4200 TapeStation system (Agilent Technologies, Santa Clara, CA, USA).

ONT long-read sequencing was performed using the SQK-LSK114 kit and three FLO-PRO114M flow cells on the P2-Solo platform, yielding 87.7 Gbp of raw reads, with a mean read size of 5.91 Kbp and read N50 of 10.48 Kbp. All ONT data were basecalled using the dna_r10.4.1_e8.2_400bps_sup@v4.2.0 model. Reads shorter than 5 Kbp or with a Q score below 7 were filtered out. To improve read quality, an error correction step was performed using Haplotype-aware ERRor cOrrection (HERRO)[13] tool, resulting in 45 Gbp of corrected reads. Only reads longer than 30 Kbp were retained for further analyses.

For PacBio sequencing and High-throughput Chromosome conformation capture (Hi-C), leaf tissue samples were sent to the Arizona Genomics Institute, where nuclei isolation, DNA and HMW DNA extraction were done. Two SMRTbell libraries were constructed with the SMRTbell Prep Kit 3.0 (Pacific Biosciences, CA, USA). Sequencing was performed on the REVIO platform at AGI, generating 234.64 Gbp of HiFi reads with a mean read size of 17 Kbp and read N50 of 18 Kbp, to ensure high-quality genome assembly. The Hi-C library was performed using the High Coverage Hi-C kit (Arima Genomics), which proprietary restriction enzymes cuts at ^GATC, G^ANTC, C^TNAG, and T^TAA sites. Hi-C library sequencing was conducted by Arima Genomics on Illumina NovaSeq X Plus platform, generating 69.6 Gbp of raw data of 2x150bp (461.4 million reads).

De novo genome assembly and chromosome level scaffolding

Genome features were estimated with GenomeScope2[14], by computing a k-mer spectra using Meryl[15], and on the PacBio HiFi reads. *A. aculeata* is a diploid species (2n = 30) with an estimated genome size of 2,84 Gbp, based on the average 2C value of 5,81 pg[16].

However, GenomeScope2 results estimates an monoploid genome size of 1.86 Gbp (Supplementary figure 1), 38.4% of non-repetitive fraction of the genome, and a heterozygosity among the two haplophases of the genome 0.617%.

A draft diploid genome assembly was generated using Hifiasm[17] v0.24.0-r702 integrating all 30+ Kbp ONT, PacBio HiFi and Hi-C data, applying default parameters. This assembly produced 683 primary contigs with a total length of 2.04 Gbp and an N50 of 141.4 Mbp. Haplotigs from the primary assembly were subsequently removed using purge_dups[18] v1.2.5, yielding a final draft assembly of 23 contigs with a total length of 1.96 Gbp.

Scaffolding was performed using the purged draft genome and the Hi-C data using the Juicer v1.6 pipeline[19] with assembly mode, following the instructions provided by Arima Genomics[20] for the high-coverage kit. Re-scaffolding and interactive Hi-C contact map were generated with YaHS[21] v1.2.2. Manual curation of contigs and scaffolds was conducted using Juicebox[22] v2.17, resulting in 27 curated contigs organized into 15 pseudochromosomes, containing 7 gaps.

Gene prediction and Functional annotation

De novo repeat families were identified and classified using RepeatModeler[23] v2.0.6 with the -LTRStruct option, resulting in the construction of a custom repeat library. Subsequently, the Viridiplantae dfam database was retrieved and combined with the *de novo* library for masking using RepeatMasker[24] v4.1.8. We identified 1.51 Gbp of interspersed repeat regions, comprising 77.36% of the genome (Table 1). Among these regions, LTR elements represent the majority, composed of 54.98%, where 35.82% being Ty1/Copia and 14.91%, Gypsy/DIRS1.

	Number of elements	Length occupied	Percentage of sequence
Retroelements	814456	1095095850 bp	55.75 %
SINEs	1	103 bp	0.00 %
Penelope	6142	669514 bp	0.03 %
LINEs	23309	14487163 bp	0.74 %
L2/CR1/Rex	2	111 bp	0.00 %

RTE/Bov-B	4579	1345562 bp	0.07 %
L1/CIN4	18728	13141490 bp	0.67 %
LTR elements	785004	1079939070 bp	54.98 %
Ty1/Copia	400318	703566069 bp	35.82 %
Gypsy/DIRS1	194020	292895927 bp	14.91 %
DNA transposons	50091	52004008 bp	2.65 %
hobo-Activator	7164	6177870 bp	0.31 %
MULE-MuDR	10858	6602619 bp	0.34 %
Tourist/Harbinger	2	117 bp	0.00 %
Rolling-circles	2335	2498242 bp	0.13 %
Unclassified	1003624	372440679 bp	18.96 %
Total interspersed repeats	-	1519540537 bp	77.36 %
Small RNA	2766	5711707 bp	0.29 %
Simple repeats	256716	13560740 bp	0.69 %
Low complexity	40845	2153093 bp	0.11 %

Table 1. Annotation of repeat content for *Acrocomia aculeata*.

Using the chromosome level assembly, gene prediction was performed incorporating RNA-seq datasets from 21 accessions, comprising different tissues of the palm tree (Table 2) - including fruit pulp, male and female flowers, roots, and leaves (NCBI BioProject PRJNA489676[25]). Each dataset was aligned to the 15 pseudochromosomes using HISAT2[26] v2.2.1 with --rna-strandness RF and --dta, and the resulting alignment files were merged with samtools[27] v1.21.

After processing each RNA-seq data, we analyzed the mappability percentage for each tissue. The percentage of reads aligned to the assembled genome varied from 36.21% to 92.47%, with an average of 78.79%. Leaf tissues, female flowers and bulb were the plant parts that got the highest mapped reads, indicating a high level of completeness and structural accuracy in genic regions, allowing effective transcriptome-supported annotation.

Accession number	Sample part	Base pairs	Read number	Mappability (%)
SRR7822563	Leaf	4,2 Gbp	20938255	92,47
SRR7822564	Leaf Sheath	5,9 Gbp	29394232	91,79
SRR7822565	Female Flower	15,7 Gbp	78987778	81,24
SRR7822566	Female Flower	5,5 Gbp	27824979	82,24
SRR7822567	Leaf	4,8 Gbp	24284522	92,39
SRR7822568	Leaf	4,6 Gbp	23053873	92,54
SRR7822569	Bulb	4,1 Gbp	20773350	91,67
SRR7822570	Bulb	5,1 Gbp	25893211	92,60
SRR7822571	Bulb	4,6 Gbp	23313242	91,47
SRR7822572	Female Flower	5,7 Gbp	28635910	54,10
SRR7822573	Root	4,9 Gbp	24696967	83,62
SRR7822574	Root	4,3 Gbp	21595079	84,03
SRR7822575	Root	4,7 Gbp	23561512	86,58
SRR7822576	Leaf Sheath	5,5 Gbp	27501990	93,17
SRR7822577	Leaf Sheath	5,1 Gbp	25450315	92,81
SRR7822578	Male Flower	4,5 Gbp	22564627	76,61
SRR7822579	Male Flower	5 Gbp	24850587	72,59
SRR7822580	Male Flower	4,5 Gbp	22487082	85,94
SRR7822581	Ripening fruit	6,4 Gbp	31804017	32,90
SRR7822582	Ripening fruit	5,9 Gbp	29007834	47,66
SRR7822583	Ripening fruit	6,5 Gbp	31986486	36,21

Table 2. Read mappability across currently available NCBI *A. aculeata*'s RNA-seq datasets.

The quality and coverage of the aligned reads for each transcript were assessed using QualiMap[28] v2.2.2, to avoid misleading extrinsic evidence from RNA-seq quality, we filtered out datasets with read mappability lower than 60% (SRR7822572, SRR7822581, SRR7822582 and SRR7822583). The final BAM file was indexed. NCBI RefSeq protein sequences for *Elaeis guineensis* were downloaded (48.771 sequences). Both RNA-seq and

protein data were used as extrinsic evidence for gene prediction using BRAKER3[29–35] v3.0.8 with --softmasking and --gff3 parameters.

The resulting gene prediction was then filtered using AGAT[36] v1.0.0 to retain only genes with transcript support and remove conflicting features, maintaining only the longest isoform for each predicted gene and reducing the redundancy, resulting in a total of 28.204 predicted genes.

To functionally annotate the predicted proteins, we utilized EggNOG-mapper[37] v2.1.8, PANNZER2[38] and Automated Assignment of Human Readable Descriptions (AHRD) tool[39] v3.3.3 function descriptions, providing functional insights into the predicted gene set. All predicted proteins were functionally annotated by at least one source.

Identification of non-coding RNAs

We employed a combination of curated tools optimized for different RNA classes. Firstly, we used tRNAscan-SE[40] v1.4 with default parameters to locate transfer RNA (tRNA) genes. Ribosomal RNAs (rRNA) and other non-coding RNAs were predicted using the RFAM database with Infernal's cmscan[41] v1.1.5 command, filtering out predictions with E-value greater than 0.05. A total of 5761 rRNA, 1679 tRNAs, 932 snoRNAs, 162 miRNAs and 138 snRNAs were identified.

Technical Validation

Gene space completeness of the genome assembly was assessed using BUSCO[42] v5.7 with the *liliopsida_odb10* lineage, contiguity metrics were assessed with QUAST[43] v5.3.0. Base-level quality (QV), k-mer completeness and false duplication percentage were estimated with Merqury[15] v1.3 using PacBio HiFi data. Pseudochromosome structure and genomic features were visualized using a CIRCOS plot, displaying gene density, Gypsy and Copia elements densities, GC content and pseudochromosome collinearity. Collinearity analyses were performed with MCScanX[44] v1.0.0 (Figure 1).

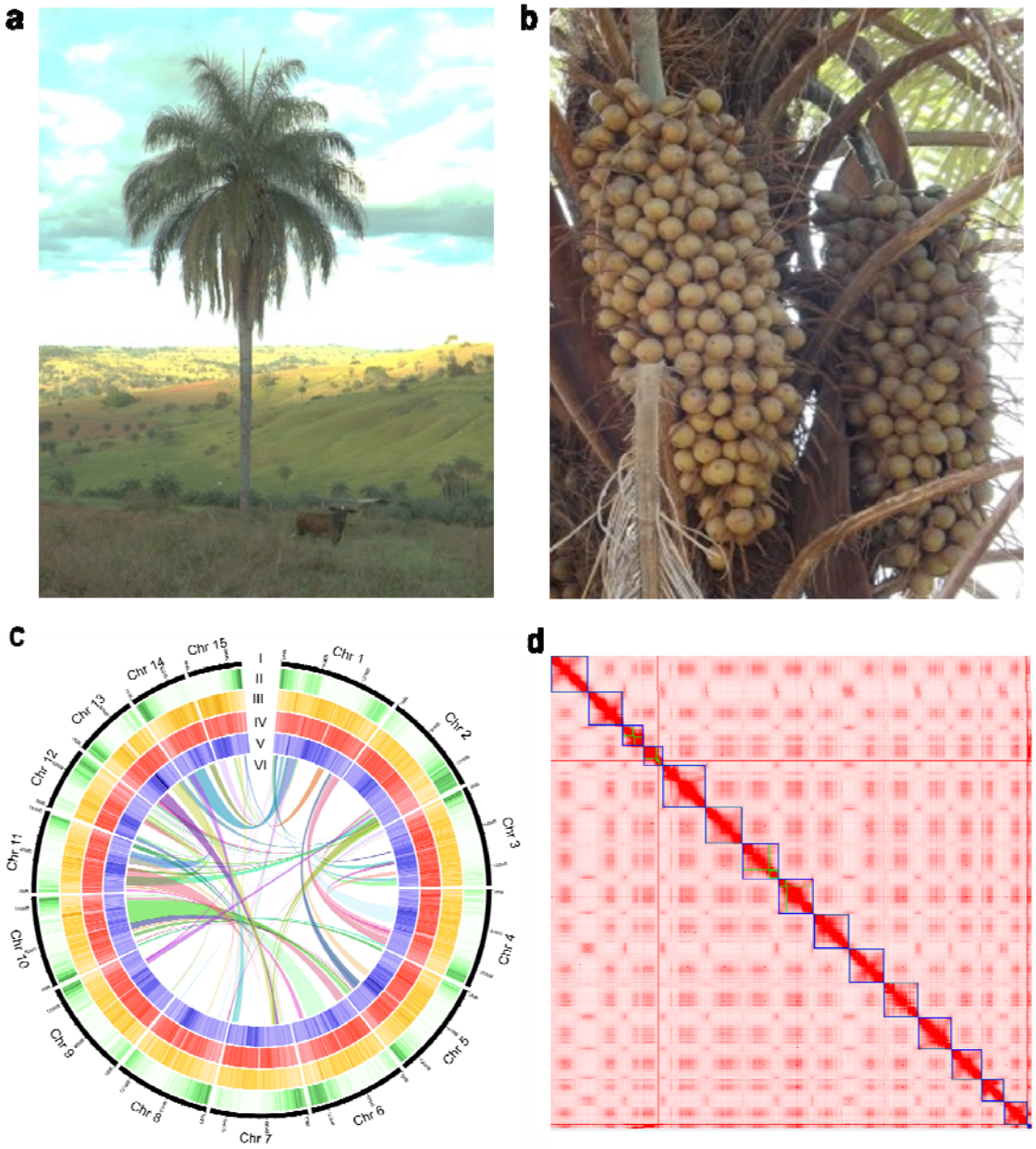


Figure 1: Morphology and genomic features of *Acrocomia aculeata*. (a) Adult individual, located at Carmo do Paranaiba, Minas Gerais, in frontal view, showing a solitary stipe and pinnately compound leaves. (b) Infructescence bearing globose drupes with a yellow epicarp. (c) Circos plot of the *A. aculeata* genome; tracks represent (I) chromosomes, (II) gene density, (III) LTR/Gypsy element density, (IV) LTR/Copia element density, (V) GC content, and (VI) chromosome collinearity blocks (self vs. self). (d) Hi-C contact map illustrating chromatin interactions across the genome. Photographs a) and b) taken by Suellen Alves Vianna.

The final assembly is highly contiguous, with a N50 of 143.43 Mbp and a scaffold assembly of all 15 pseudochromosomes covering a total of 1.94 Gbp (Table 3). The BUSCO analysis showed that our assembly has 82.3% and 14.8% of complete and duplicated

orthologues genes, with only 2.5% and 0.4% fragmented or missing. Alongside that, a low false duplication of 1.71%, K-mer completeness of 90.58% and QV greater than 75, demonstrates a high-quality assembly. Gene prediction with both RNA-seq and *E. guineensis*' proteins, has a BUSCO completeness of 98.3%, with 83.0% single copy and 15.3% duplicated.

Assembly	Genome length	Contigs number	N50	Assembly BUSCO	Proteins BUSCO	QV	K-mer completeness	False duplication
contig assembly	1.96 Gbp	27	143.43 Mbp	97.1%	-	75.66	90.69%	1.72%
scaffold assembly	1.94 Gbp	15	143.43 Mbp	97.1%	98.3%	76.40	90.58%	1.71%

Table 3. Summary of genome metrics for *Acrocomia aculeata*.

The Telomere-to-Telomere toolkit[45] v1.2.5 (quarTeT) and Telomere Identification ToolKit[46] v0.2.63 (TIDK) were employed to identify and create the plot to visualize telomeric repeats across all pseudochromosomes, resulting in the sequence AAACCCT. We assembled nine pseudochromosomes with telomeric repeats on both ends, and the remaining six pseudochromosomes with a telomeric repeat at only one end (Supplementary figure 2 and 3).

We compared the 15 peudochromosomes of *A. aculeata* against the genomes of four palm species (Figure 2): *Elaeis guineensis*, *E. oleifera*[47], *Cocos nucifera*[48] and *Phoenix dactylifera*[49], using D-Genies web service[50]. These results show higher levels of collinearity between *A. aculeata* and *E. guineensis*, *E. oleifera* and *C. nucifera* than with *P. dactylifera* which is expected, because these palm species are from the same subfamily Cocosaea. *E. guineensis*, *E. oleifera* and *A. aculeata* diverged around 30 Ma. *C. nucifera*'s subtribe diverged at the end of Palaeogeno, around 60 Ma. Finally, for *P. dactylifera* which are from Coryphoideae, there are more differences between chromosomes, containing fragmented collinearity regions[51].

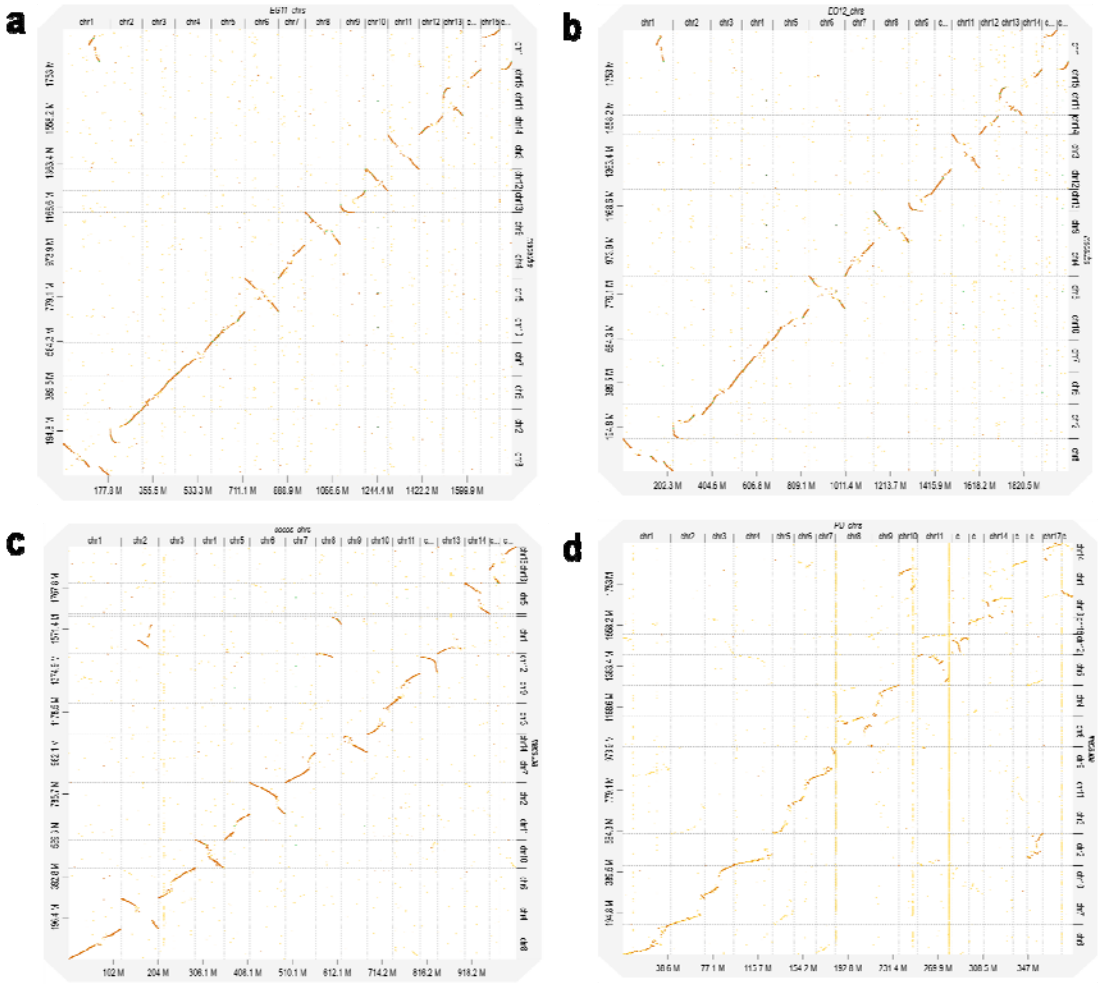


Figure 2: Base-by-base comparison between *Acrocomia aculeata* and other palm species. (a) *Elaeis guineensis* (African oil palm; EG11; PRJNA192219). (b) *Elaeis oleifera* (American oil palm; EO12.1; PRJNA183707). (c) *Cocos nucifera* (Coconut palm; Cnuc_CPCRI_v2; PRJNA413280). (d) *Phoenix dactylifera* (Date palm; Barhee BC4; PRJNA322046).

It is possible to observe some chromosome rearrangements (highlighted in red). *E. guineensis* and *A. aculeata* shows inverted regions between chromosome 2 and 8 and a spliced region shared with chromosome 1, also having inverted collinearity between pseudochromosomes 3, 5, 9 and 11 and chromosomes 11, 6, 8 and 13. *Elaeis oleifera* exhibits a collinearity pattern largely congruent with *E. guineensis*, but with distinct differences in each chromosome, such as chromosomes 1 and 2. *C. nucifera* presents less similar regions than *E. guineensis* and *E. oleifera*, with more rearrangements, found at pseudochromosomes 1, 2, 4, 5, 10 and 13. Pseudochromosome 1 also shows fragmentation between different *C. nucifera* chromosomes 2 and 8. Finally, the comparison between *P. dactylifera* reinforces the

phylogenetic distance between the species, with high fragmented regions such those found at pseudochromosomes 1, 2, 4, 5, 6, 12, 13 and 14.

Code availability

We used publicly available programs with default parameters; different options were described in the Methods.

References

1. Harri Lorenzi, N., Larry Kahn, Francis Ferreira, Evandro José Linhares. *Flora Brasilieira: Arecaceae (Palmeiras)*. (Plantarum, 2010).
2. Vianna, S.A, C.-R. Acrocomia in Flora e Funga do Brasil.
3. Barcelos, E. *et al.* Oil palm natural diversity and the potential for yield improvement. *Front. Plant Sci.* **6**, (2015).
4. Colombo, C. A., Chorfi Berton, L. H., Diaz, B. G. & Ferrari, R. A. Macauba: a promising tropical palm for the production of vegetable oil. *OCL* **25**, D108 (2018).
5. Lopes, D. D. C., Steidle Neto, A. J., Mendes, A. A. & Pereira, D. T. V. Economic feasibility of biodiesel production from Macauba in Brazil. *Energy Econ.* **40**, 819–824 (2013).
6. Janssen, R. & Rutz, D. D. Sustainability of biofuels in Latin America: Risks and opportunities. *Energy Policy* **39**, 5717–5725 (2011).
7. Coser, S. M., Motoike, S. Y., Corrêa, T. R., Pires, T. P. & Resende, M. D. V. Breeding of Acrocomia aculeata using genetic diversity parameters and correlations to select accessions based on vegetative, phenological, and reproductive characteristics. *Genet. Mol. Res.* **15**, (2016).
8. Lanes, É. C. M., Motoike, S. Y., Kuki, K. N., Resende, M. D. V. & Caixeta, E. T. Mating System and Genetic Composition of the Macaw Palm (*Acrocomia aculeata*): Implications for Breeding and Genetic Conservation Programs. *J. Hered.* **107**, 527–536 (2016).

9. Simiqueli, G. F., Resende, M. D. V. D., Motoike, S. Y. & Henriques, E. Inbreeding depression as a cause of fruit abortion in structured populations of macaw palm (Acrocomia aculeata): Implications for breeding programs. *Ind. Crops Prod.* **112**, 652–659 (2018).
10. Díaz-Hernández, B. G. *et al.* Assessing the genetic vulnerability of MACAÚBA palm [*Acrocomia aculeata* (JACQ.) LODD. ex MART.] through the mating system and genetic diversity of open-pollinated progenies. *Ann. Appl. Biol.* **184**, 238–249 (2024).
11. Martins, K. K. M. *et al.* Neotropical palms: from their conservation to economic potential. *Front. Plant Sci.* **15**, 1487297 (2024).
12. Jones, A. *et al.* Scalable high-molecular weight DNA extraction for long-read sequencing v1. Preprint at https://doi.org/10.17504/protocols.io.bnjhmcj6 (2020).
13. Stanojevic, D., Lin, D., Nurk, S., Florez De Sessions, P. & Sikic, M. Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. Preprint at https://doi.org/10.1101/2024.05.18.594796 (2024).
14. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
15. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
16. Abreu, I. S., Carvalho, C. R., Carvalho, G. M. A. & Motoike, S. Y. First karyotype, DNA C-value and AT/GC base composition of macaw palm (Acrocomia aculeata, Arecaceae) - a promising plant for biodiesel production. *Aust. J. Bot.* **59**, 149 (2011).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
18. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

19. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
20. Arima Genomics, A. G. Bioinformatics-User-Guide-Arima-HiC-and-Arima-High-Coverage-HiC.
21. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
22. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
23. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
24. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. (2015).
25. Bazzo, B. R., De Carvalho, L. M., Carazzolle, M. F., Pereira, G. A. G. & Colombo, C. A. Development of novel EST-SSR markers in the macaúba palm (Acrocomia aculeata) using transcriptome sequencing and cross-species transferability in Arecaceae species. *BMC Plant Biol.* **18**, 276 (2018).
26. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
27. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
28. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
29. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47**, (2014).
30. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).

31. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
32. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
33. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
34. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
35. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644 (2008).
36. Jacques Dainat *et al.* NBISweden/AGAT: AGAT-v1.4.1. Zenodo https://doi.org/10.5281/ZENODO.3552717 (2024).
37. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
38. Törönen, P., Medlar, A. & Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).
39. Florian Boecker. AHRD: Automatically Annotate Proteins with Human Readable Descriptions and Gene Ontology Terms. (Rheinische Friedrich-Wilhelms-Universität Bonn, 2021).
40. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686–W689 (2005).
41. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

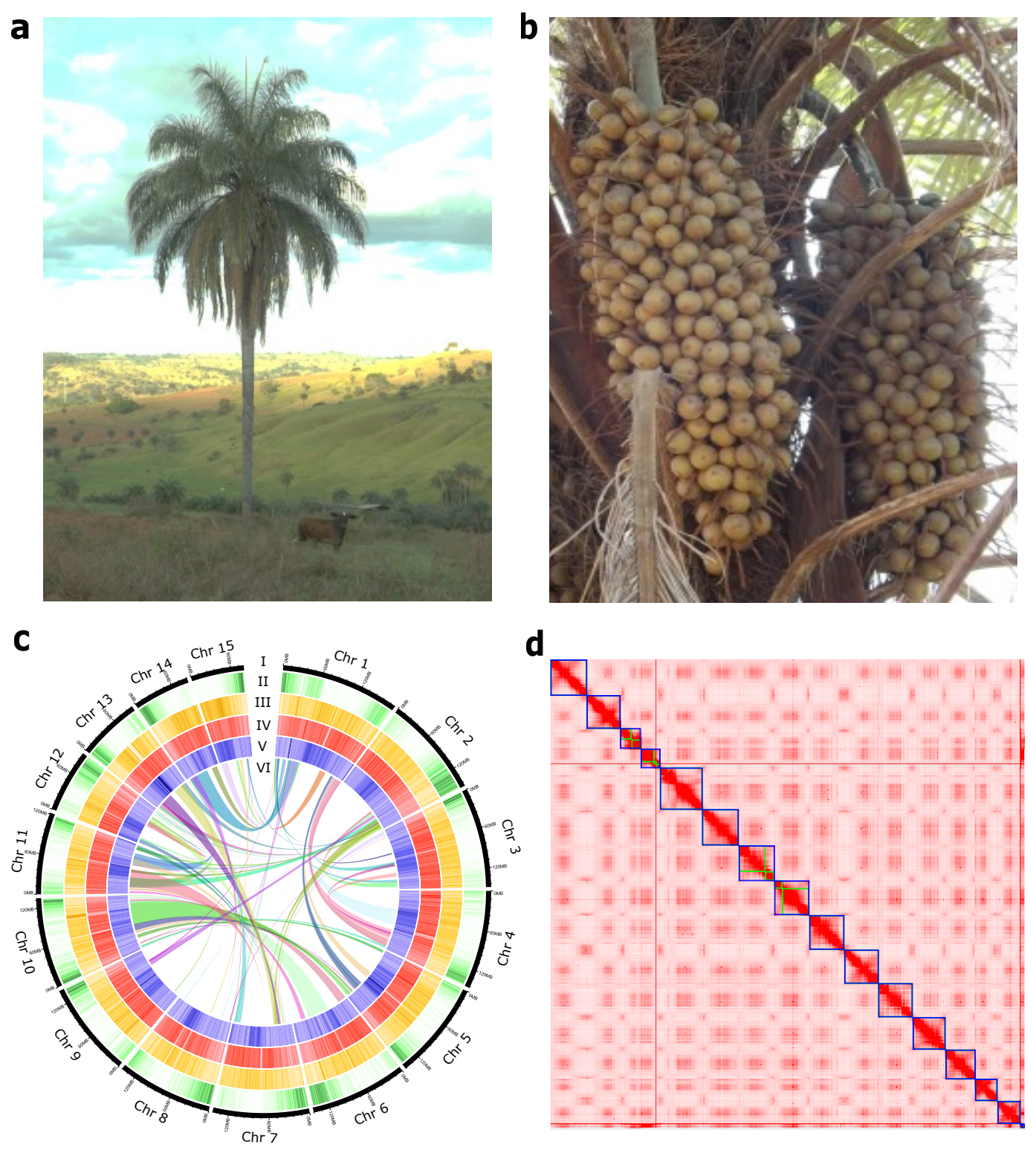
42. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
43. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
44. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
45. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
46. Brown, M. R., Manuel Gonzalez De La Rosa, P. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, btaf049 (2025).
47. Low, E.-T. L. *et al.* Chromosome-scale *Elaeis guineensis* and *E. oleifera* assemblies: comparative genomics of oil palm and other Arecaceae. *G3 Genes Genomes Genet.* **14**, jkae135 (2024).
48. Muliyar, R. K. *et al.* Assembly and Annotation of the Nuclear and Organellar Genomes of a Dwarf Coconut (Chowghat Green Dwarf) Possessing Enhanced Disease Resistance. *OMICS J. Integr. Biol.* **24**, 726–742 (2020).
49. Hazzouri, K. M. *et al.* Genome-wide association mapping of date palm fruit traits. *Nat. Commun.* **10**, 4680 (2019).
50. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
51. Baker, W. J. & Couvreur, T. L. P. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical biogeography. *J. Biogeogr.* **40**, 274–285 (2013).

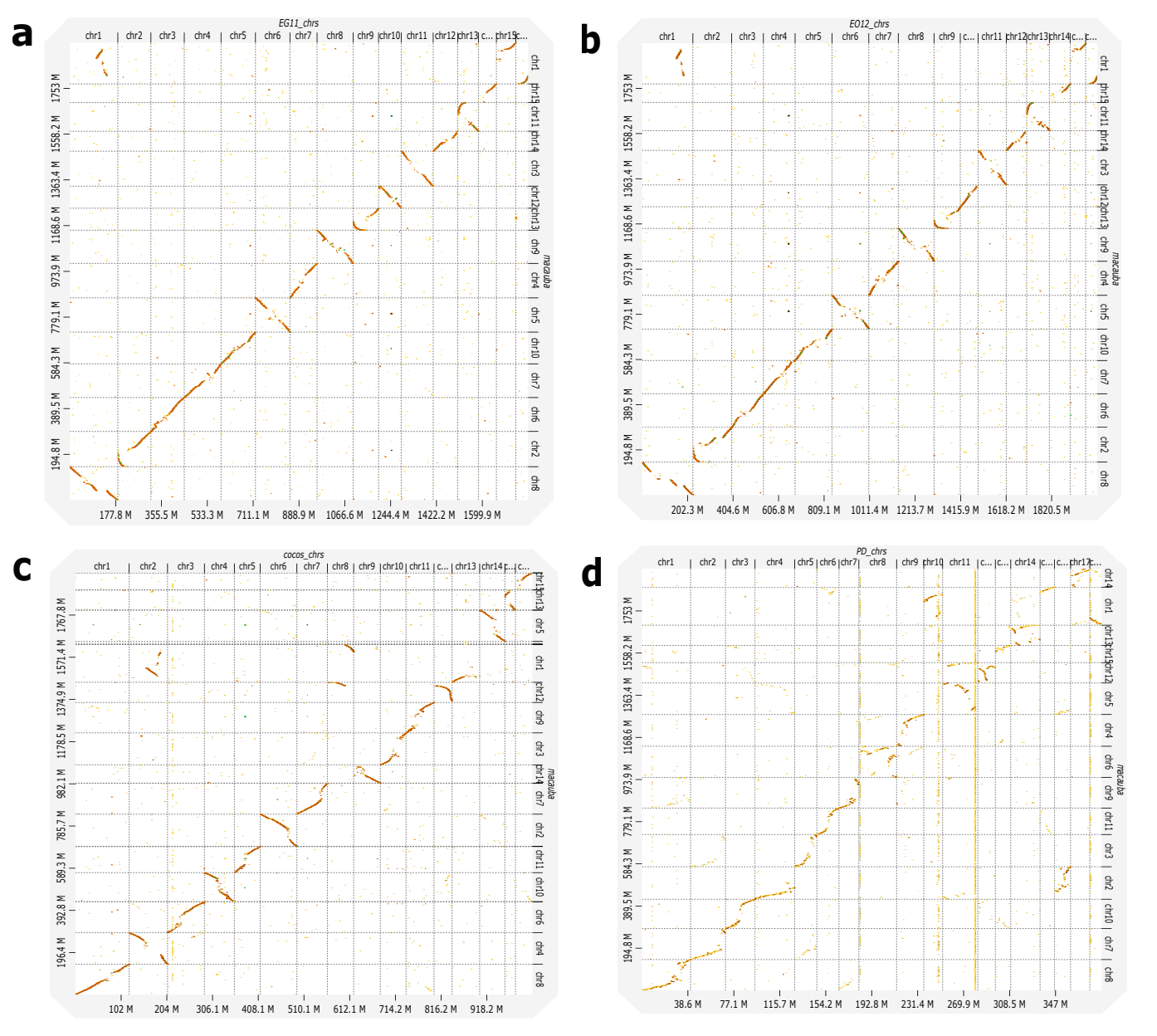
Acknowledgements

Acelen Renovaveis supported this work by funding the costs of the PacBio HiFi and Hi-C sequencing data (Project number: FEALQ 104.781). FAPESP 2021/10319-0 and CNPq 313417/2023-7 also supported this work for all ONT sequencing data.

Author contributions

M.S. and M.I.Z. conceptualized and designed this study and manuscript review. C.B.G., and M.S, HMW DNA extraction, library preparation, Nanopore DNA sequencing and manuscript preparation. C.B.G., I.A. and J.R. CTAB DNA extraction and Illumina library preparation., M.I.Z., D.D.N., C.A.L., J.B.P. funding coordination, C.C. sample collection. M.S. technical validation, bioinformatic analysis and data submission. D.M.R.P., R.V.S. and J.M.M. manuscript review.


a
b
c
d
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
I
II
III
IV
V
VI


a
EG11_chrs
macauba
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 c... chr19 c...
chr1 chr19 chr11 chr14 chr3 chr12 chr13 chr9 chr4 chr5 chr10 chr7 chr6 chr2 chr8
1753 M 1558.2 M 1363.4 M 1168.6 M 973.9 M 779.1 M 584.3 M 389.5 M 194.8 M
177.8 M 355.5 M 533.3 M 711.1 M 888.9 M 1066.6 M 1244.4 M 1422.2 M 1599.9 M
b
EO12_chrs
macauba
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 c... chr11 chr12 chr13 chr14 c... c...
chr1 chr19 chr11 chr14 chr3 chr12 chr13 chr9 chr4 chr5 chr10 chr7 chr6 chr2 chr8
1753 M 1558.2 M 1363.4 M 1168.6 M 973.9 M 779.1 M 584.3 M 389.5 M 194.8 M
202.3 M 404.6 M 606.8 M 809.1 M 1011.4 M 1213.7 M 1415.9 M 1618.2 M 1820.5 M
c
cocos_chrs
macauba
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 c... chr13 chr14 c... c...
chr13 chr5 chr1 chr12 chr9 chr3 chr14 chr7 chr2 chr11 chr10 chr6 chr4 chr8
1767.8 M 1571.4 M 1374.9 M 1178.5 M 982.1 M 785.7 M 589.3 M 392.8 M 196.4 M
102 M 204 M 306.1 M 408.1 M 510.1 M 612.1 M 714.2 M 816.2 M 918.2 M
d
PD_chrs
macauba
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 c... c... chr14 c... c... chr17 c...
chr14 chr1 chr13 chr15 chr12 chr5 chr4 chr6 chr9 chr11 chr3 chr2 chr10 chr7 chr8
1753 M 1558.2 M 1363.4 M 1168.6 M 973.9 M 779.1 M 584.3 M 389.5 M 194.8 M
38.6 M 77.1 M 115.7 M 154.2 M 192.8 M 231.4 M 269.9 M 308.5 M 347 M