Repetita iuvant. The extraordinary evolutionary history of the endemic palm *Chamaerops humilis* L. in the Mediterranean

Guarino Carmine[1a], Labella-Ortega Mónica[1*a], Cipriani Guido[2], Schicchi Rosario[3], Bazan Giuseppe[4], Hengchi Chen[5,6], Ranauda Maria Antonietta[1], Maisto Maria[1], Zuzolo Daniela[1], Tartaglia Maria[1], Prigioniero Antonello[1]

[1] Department of Science and Technology, University of Sannio, via de Sanctis snc, Benevento, 82100, Italy

[2] Department of Agriculture, Food, Environmental and Animal Science, University of Udine, Via delle Scienze 206, 33100 Udine, Italy

[3] Department of Agricultural, Food and Forestry Sciences, University of Palermo, Viale Delle Scienze, Ed. 4, 90128, Palermo, Italy

[4] Department of Biological, Chemical and Pharmaceutical Sciences and Technologies (STEBICEF), Università degli Studi di Palermo, Viale delle Scienze, ed. 17, 90128 Palermo, Italy

[5] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent 9052, Belgium

[6] VIB Center for Plant Systems Biology, VIB, Ghent 9052, Belgium

* Author for correspondence: mlabellaortega@unisannio.it (M. Labella-Ortega)

[a] Equal contribution

Supplementary Information

Supplementary results

Genome assembly

For the de novo assembly of the *C. humilis* genome, three sequencing and assembly technologies were used. Initially, the kmer-based analysis estimated a haploid genome size of approximately 2.2 Gbp, a low heterozygosity rate of 2.3%, and a high repeat content exceeding 70%. Using PacBio reads, a total of 10,638 contigs were assembled, achieving a contig N50 value of 1,023,919 bp, an estimated genome size of 3.49 Gb and a coverage of 58.1X. In the primary contig assembly, BUSCO analysis predicted that single-copy genes exhibit a completeness of 84%, with a low percentage of duplicated (9.2%), fragmented (3.1%), and missing (12.9%) genes.

A polishing step was undertaken to refine the genome assembly using Illumina short reads, which provide a coverage of 38.9X. This step corrected approximately 20.5 Mbp (1% of the genome). Following the mapping of the short reads to the draft genome, de novo assembly of the unmapped reads resulted in 7,553 new contigs, totalling 30,280,817 bp, with an N50 of 4,393 and an average size of 4,009 bp. The BUSCO analysis revealed a significant increase in the percentage of completed genes (91%) and a slight reduction in fragmented (2.6%) and missing genes (12.7%). In the redundancy evaluation, PacBio reads mapped back onto the raw genome assembly contigs showed one clear main peak, highlighting a low fraction of redundant or secondary contigs in the assembly (**Supplementary Fig. 8**).

To construct the scaffold-scale assembly, Hi-C data was used, with a Hi-C coverage of 18.6X. This approach resulted in 250 scaffolds obtained by merging 13,313 contigs. The remaining 4,514 unscaffolded contigs, representing 0.5% of the genome, were merged into an artificial scaffold.

The 251 scaffolds of our genome assembly were mapped against the genome from NCBI (GCA_042465325.1) and assigned/oriented to each chromosome. 240 of the scaffolds were assembled into 18 chromosomes, and the remaining 11 were assembled into an artificial chromosome (chromosome 0) and arbitrary gaps of 100 bp were added between consecutive contigs (**Fig. 1, Supplementary Fig. 9**).

The final high-quality assembled genome consisted of 3.44 Gbp, being more less, approximately 3.6 times as larger as *Phoenix dactylifera* [1] and 2 times as larger as *Elaeis guineensis* [2]. The longest chromosome was the Chr2 with 345.89 Mbp, and the shortest was the Chr18 with 65.48 Mbp (**Supplementary Table 3**).

Supplementary Methods

Plant materials and nucleic acid extraction

The plant material from which nucleic acid was extracted consisted of young leaves taken from the yearlings of male and female individuals of *C. humilis*. Fresh leaves were sampled from individuals living in the Circeo National Park (Latina, Italy).

DNA extraction was performed following the protocol described by Russo et al[3], with slight modifications as outlined below. Young leaves were collected, flash-frozen in liquid nitrogen, and stored at −80°C until DNA extraction. A 100 mg aliquot of frozen leaf tissue was ground to a fine powder using a mortar and pestle in liquid nitrogen. The powder was then transferred to a sterile 2 mL centrifuge tube containing 600 µL of SDS lysis buffer (1% polyvinylpyrrolidone 40 (PVP40), 1% sodium metabisulphite, 0.5 M sodium chloride, 100 mM Tris–HCl (pH 8), 50 mM EDTA (pH 8), 2% β-mercaptoethanol (β-ME), and 1.5% sodium dodecyl sulfate (SDS)). The sample was vortexed for 3–5 s and incubated in a thermomixer with gentle agitation (400 rpm for 20 minutes at 55°C), adding 4 µL of 100 mg/mL DNase-free RNase A (Qiagen, Germantown, MD, USA). After the incubation, 200 µL of 5 M potassium acetate was added, and the mixture was inverted for mixing. Next, 800 µL of phenol:chloroform:isoamyl alcohol (25:24:1, v/v, pH 8) was added, and the sample was incubated for 10 min at room temperature (RT). The sample was then centrifuged for 10 min at 10,000 × g at RT. The supernatant was carefully transferred to a new 2 mL tube, and the extraction was repeated one more time. The resulting sample was purified using magnetic beads (Sera-Mag SpeedBeads™ Carboxyl Magnetic Beads, GE Healthcare 65152105050250, Fisher Scientific). The sample underwent three washes with 1 mL of 70% ethanol (EtOH) before being eluted in 50 µL of 10 mM Tris–HCl (pH 8.5). Quality control was performed on the extracted DNA before proceeding with library preparation and sequencing.

Crosslinking of the solid tissue samples was done for HiC [4]. Briefly, 2 g of young male leaf tissue was ground into a fine powder in a liquid nitrogen-chilled mortar, once the tissue was powdered, it was resuspended in 10 times its volume of 1% formaldehyde (37%). After resuspension, the sample was incubated at room temperature for 20 min, with periodic mixing or vortexing to ensure thorough mixing. Following the incubation, glycine was added to the sample to achieve a final concentration of 125 mM, and the sample was incubated at room temperature for an additional 15 min, with periodic mixing. After the second incubation, the sample was centrifuged (1000 × g for 1 min) to pellet the crosslinked tissue powder. The pellet was then stored at −80°C until further analysis.

To aid in gene annotation and phylogenomic analyses, fresh leaves from male and female individuals of *C. humilis* were collected for RNA sequencing (RNA-seq). RNA was extracted from leaf samples, which were collected and immediately frozen in liquid nitrogen and transported to the laboratory without interrupting the cold chain. The samples were pulverized using a mortar and pestle in liquid nitrogen, and 50 mg of each sample was taken for subsequent extractions using the RNeasy Plant Mini Kit (QIAGEN) following the manufacturer's instructions. High quality RNA-seq libraries were prepared and sequenced with an Illumina NovaSeq 6000 platform.

Genome sequencing and assembly

Genomic DNA from male and female individuals of *C. humilis* was utilized to construct PacBio and Illumina sequencing libraries, while DNA from male individuals was employed for Hi-C library preparation. A SMRTbell library was constructed with Sequel 1.0 reagents using PacBio Sequel, and paired end libraries for Illumina were prepared following the manufacturer's protocol. Hi-C library

construction was performed with the Proximo™ Hi-C Plant kit (Phase Genomics) following the manufacture's protocol. Raw reads were evaluated for quality and filtered using Filtlong (v0.2.0) (https://github.com/rrwick/Filtlong) and NextDeNovo (v2.4.0)[5] and Seqtk (v4) (https://github.com/lh3/seqtk), removing poor-quality sequences and filtering out sequences < 2.5 Kbp and > 50 Kb, which showed skewed nucleotides frequency. A kmer-based statistical approach was used to estimate the genome size, heterozygosity, and repeat content of *C. humilis*. The analysis was conducted using short reads from Hi-C with Jellyfish [6] and GenomeScope tools [7].

PacBio reads were utilized for a de novo assembly using NextdeNovo v2.5.0 [5]. The process involved two stages, an initial correction of the raw reads, followed by the assembly of a consensus sequence of the corrected reads using a string graph algorithm. The completeness of the genome assembly, generated by using primary contigs, was evaluated using BUSCO (v5.0) [8]. To improve the assembly quality, a polishing step was performed with NextPolish using long and short reads [9].

To recover the potential portion of the genome not assembled in the long-reads assembly process, all short reads were mapped against the draft genome. Approximately 95-98% of the reads were successfully mapped, and unmapped reads, accounting for about 15 Gbp of data, were then extracted and used as input for a de novo assembly approach based on Spades (v3.15.5) [10].

Contigs were analysed to detect possible separation of primary contigs from secondary/alternatives ones using Purge_haplotigs pipeline [11]. Based on the plot distribution specific cutoffs were set to identify contigs with very low or high coverage (representing <1%) which were marked and excluded from the assembly. Contigs from the sequence assemblers were linked to form scaffold-scale assembly, based on ALLHi-C pipeline [12]. The completeness of the genome assembly was evaluated using BUSCO (v5.0) [8].

Scaffolds were reordered into chromosome level using RagTag [13] using a chromosome-based genome assembly of *C. humilis* (NCBI Accession: GCA_042465385.1). Each scaffold was mapped against the chromosomes and scored for confidence according to grouping, location and orientation.

Genome annotation

For gene prediction, the whole genome assembly at chromosome level, a de novo transcriptome assembly and the corresponding protein predicted, composed of 99,980 sequences, and Illumina paired-end RNA-seq data from 3 samples were used. Prior to the further analysis, a quality check was performed on the raw sequencing data, removing low quality portions while preserving the longest high quality part of NGS reads using BBDuck, normalized with BBNorm (https://sourceforge.net/projects/bbmap/) and joined to generate a comprehensive dataset of reads which were mapped on the assembly genome using STAR [14] , the total number of mapped reads was 28.8 million, 92.5% of which mapped uniquely.

MAKER2[15] was used as the core module for structural genome annotation. For the genome annotation process, repeats were soft masked using the Red algorithm [16], masking 82.3% of the genome (2.8 Gbp). A protein database combining 121,927 sequences from transcriptome assemblies, Viridiplantae UniRef90, and Coryphoideae was constructed. The soft-masked genome, protein database, and RNA-seq were used to generate training models and a draft annotation using Braker3 [17], with additional annotations provided by Helixer pipeline [18]. Proteins from *Phoenix dactylifera* were mapped using Miniprot [19], and along with all other results, were integrated using TSEBRA [20]. TransDecoder annotated CDS sequences [21], and functional annotation was performed with Pannzer2 [22].

A high-quality non-redundant TE library was generated by the EDTA pipeline [23] and the inbuilt RepeatModeler [24] . The identification of TE was performed by using RepeatMasker (v1.332)[25] utilizing the NCBI/RMBLAST search engine (v2.6.0). LAI was determined using LTR_retriever [26]. Full-length LTR retrotransposons were annotated based on the Viridiplantae database v4.0 [27].

The genome karyotype, gene distribution, and transposon density were visualized using the Circos [28]. Figures were generated in R v.4.4.2 (R Core team, 2021) using ggplot [29], ggiraphExtra [30] packages.

References

1. Hazzouri, K. M. *et al.* Genome-wide association mapping of date palm fruit traits. *Nat. Commun.* **10**, 4680 (2019).

2. Low, E.-T. L. *et al.* Chromosome-scale Elaeis guineensis and E. oleifera assemblies: comparative genomics of oil palm and other Arecaceae. *G3 GenesGenomesGenetics* **14**, jkae135 (2024).

3. Russo, A. *et al.* Low-Input High-Molecular-Weight DNA Extraction for Long-Read Sequencing From Plants of Diverse Families. *Front. Plant Sci.* **13**, 883897 (2022).

4. Padmarasu, S., Himmelbach, A., Mascher, M. & Stein, N. In Situ Hi-C for Plants: An Improved Method to Detect Long-Range Chromatin Interactions. in *Plant Long Non-Coding RNAs: Methods and Protocols* (eds. Chekanova, J. A. & Wang, H.-L. V.) 441–472 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-9045-0_28.

5. Hu, J. *et al.* NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).

6. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

7. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).

8. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

9. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).

10. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinforma.* **70**, e102 (2020).

11. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).

12. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).

13. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).

14. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

15. Campbell, M. S. *et al.* MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiol.* **164**, 513–524 (2014).

16. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16**, 227 (2015).

17. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).

18. Ekanayake, G. *et al.* DYNAMIN-RELATED PROTEIN DRP1A functions with DRP2B in plant growth, flg22-immune responses, and endocytosis. *Plant Physiol.* **185**, 1986–2002 (2021).

19. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).

20. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).

21. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

22. Törönen, P., Medlar, A. & Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).

23. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

24. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

25. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1-4.10.14 (2009).

26. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

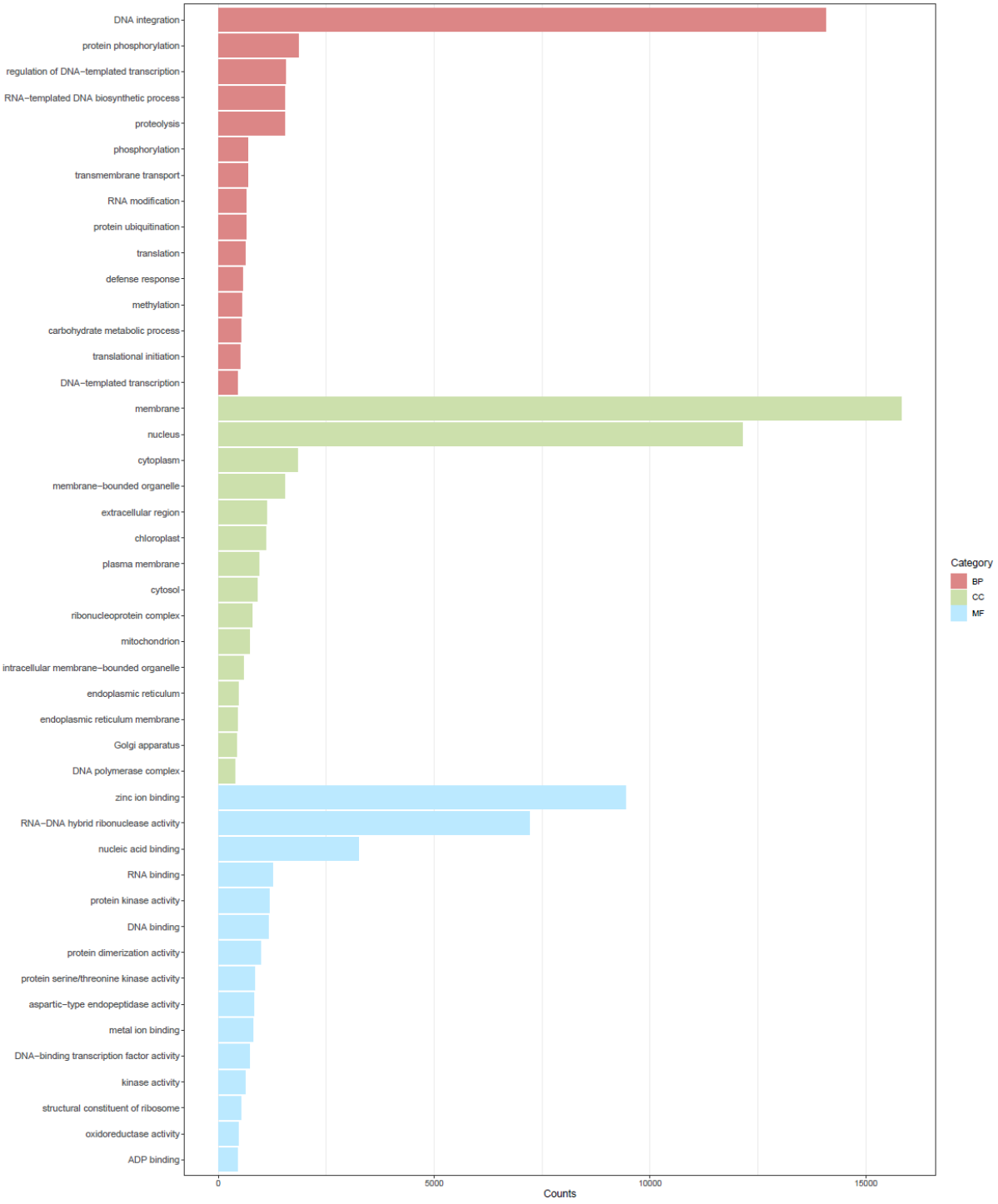
27. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, 1 (2019).

28. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

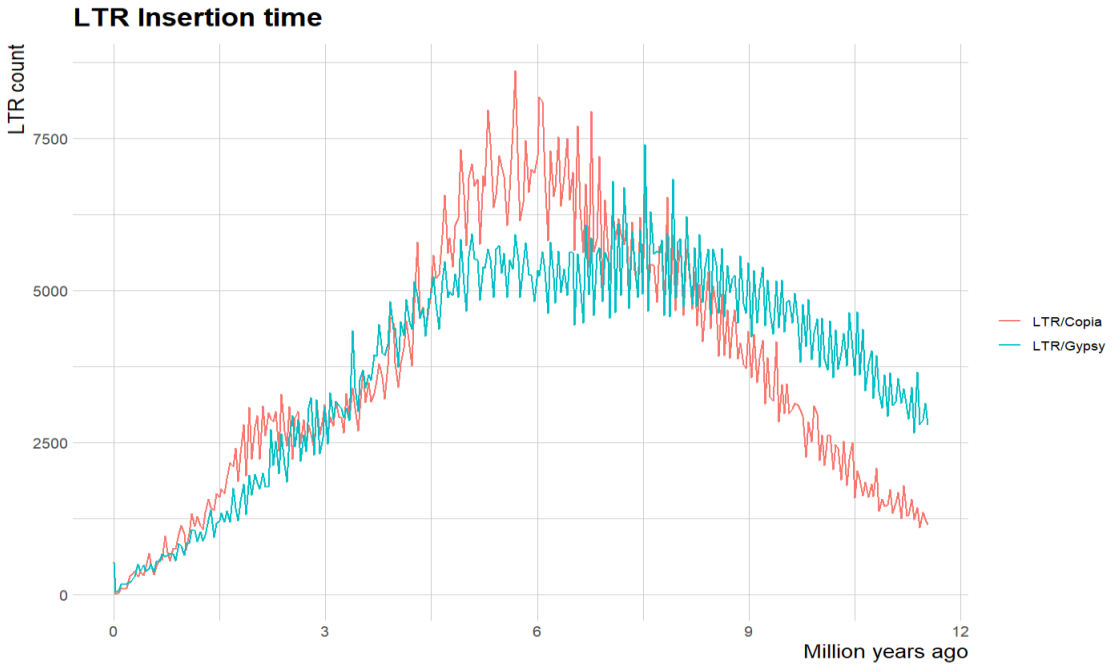
29. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* **67**, 678–679 (2011).

30. Moon, K.-W. ggiraphExtra: Make Interactive 'ggplot2'. Extension to 'ggplot2' and 'ggiraph'. 0.3.0 https://doi.org/10.32614/CRAN.package.ggiraphExtra (2016).

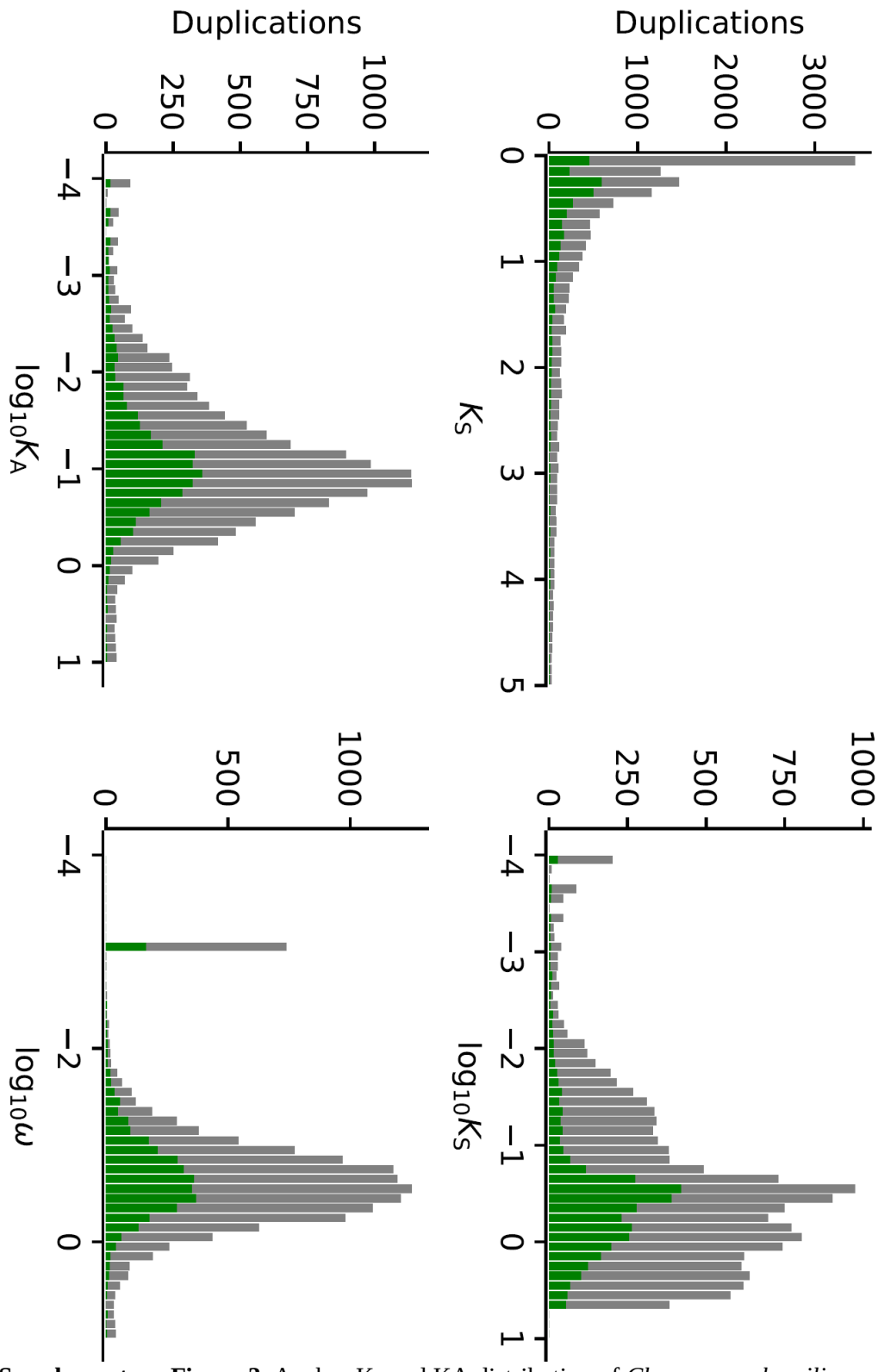
Supplementary Figure



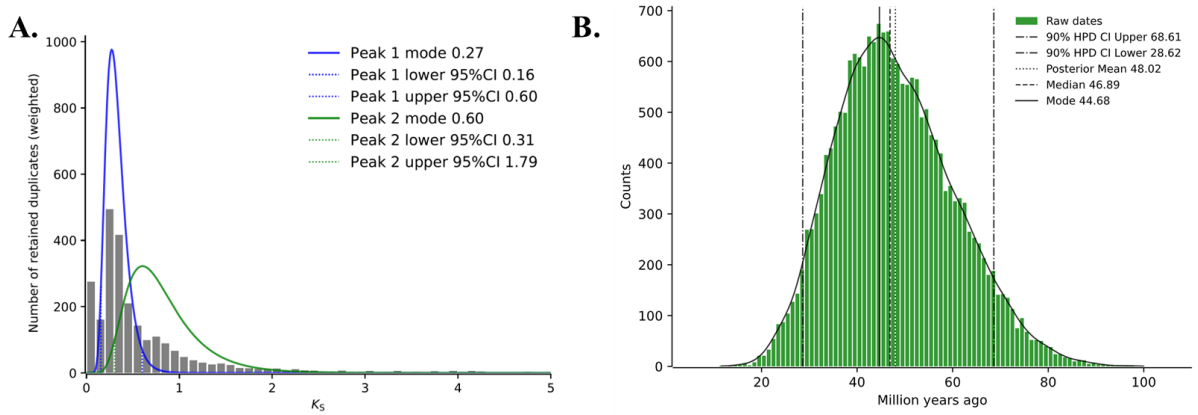
Supplementary Figure 1. GO classification of *C. humilis* genes. Distribution of GOs associated with the genes represented in the three main GO categories: biological processes (red), cellular components (green) and molecular functions (blue). The first 15 most abundant GO groups are shown.



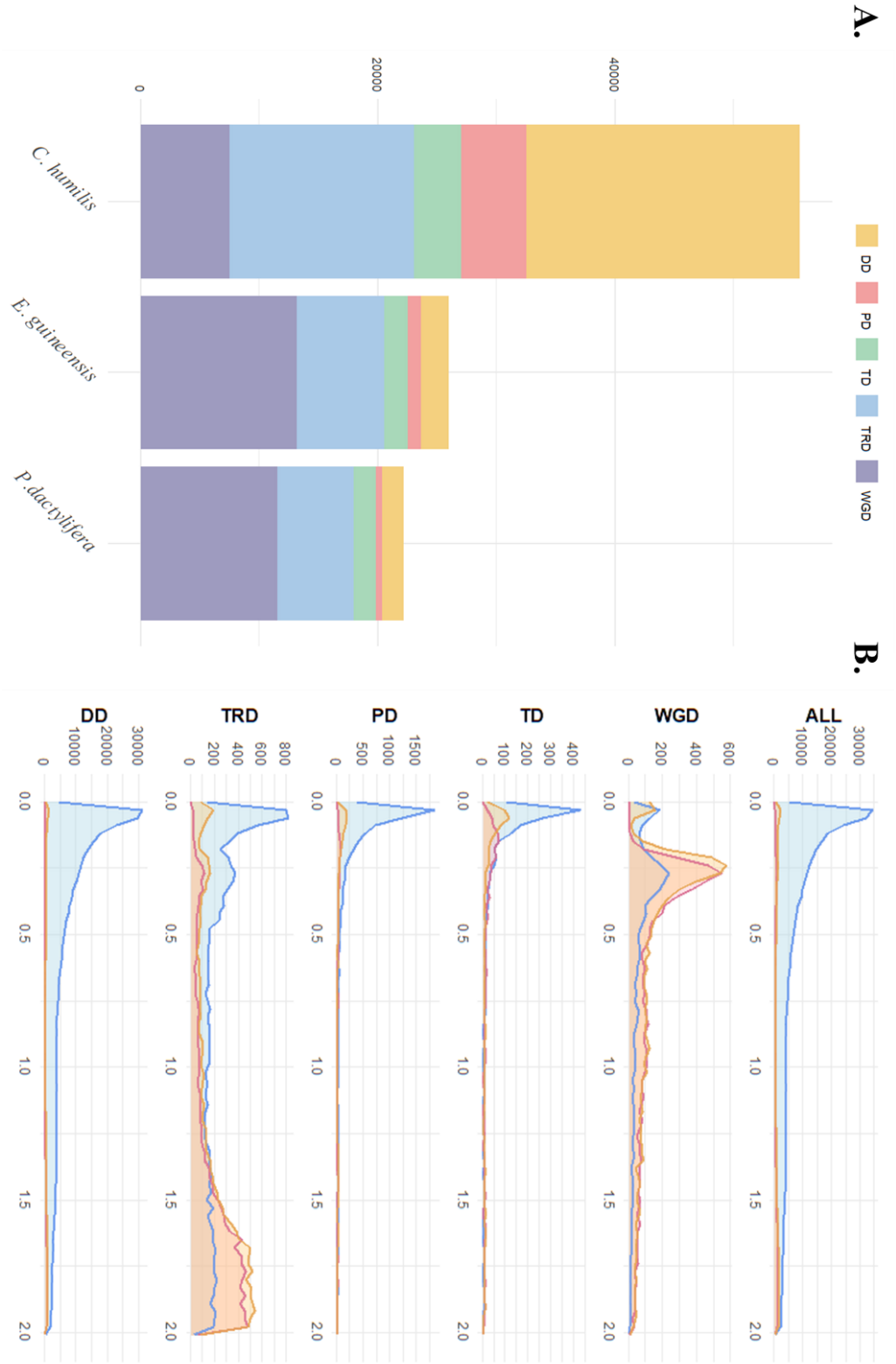
Supplementary Figure 2. Distribution of LTR transposon insertion times. Amount of LTR Gypsy (blue) and Copia (red) insertion times in million years ago.



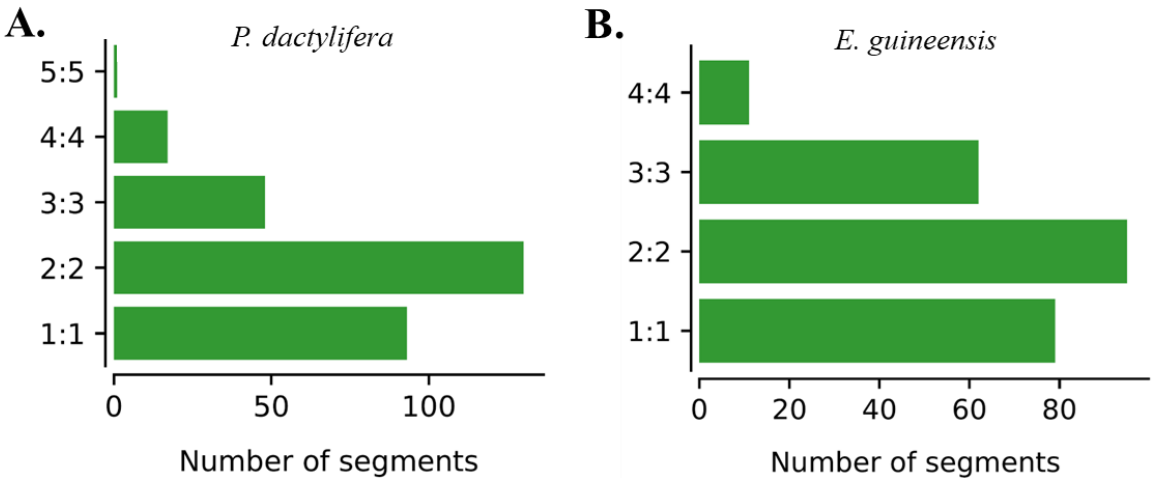
Supplementary Figure 3. Anchor Ks and KA distribution of *Chamaerops humilis*



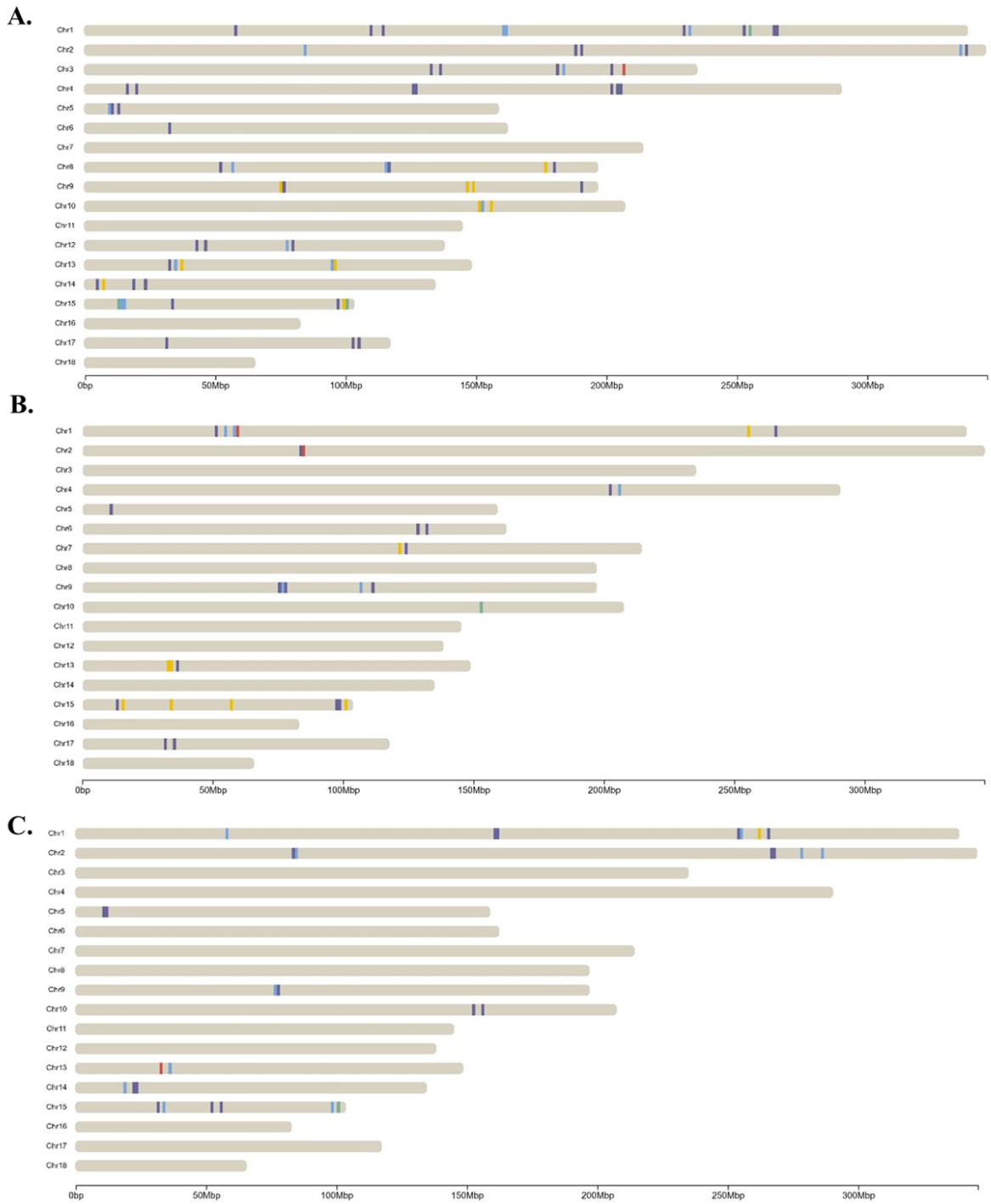
Supplementary Figure 4. Peaks emerged in the anchor pair K_S and posterior date distributions of WGD dating. **(A)** K_S age distribution for the paranome of *C. humilis* and (**B**) distribution curve, modes, overall means, peaks and 90% HPD of the date estimation of the WGD event.



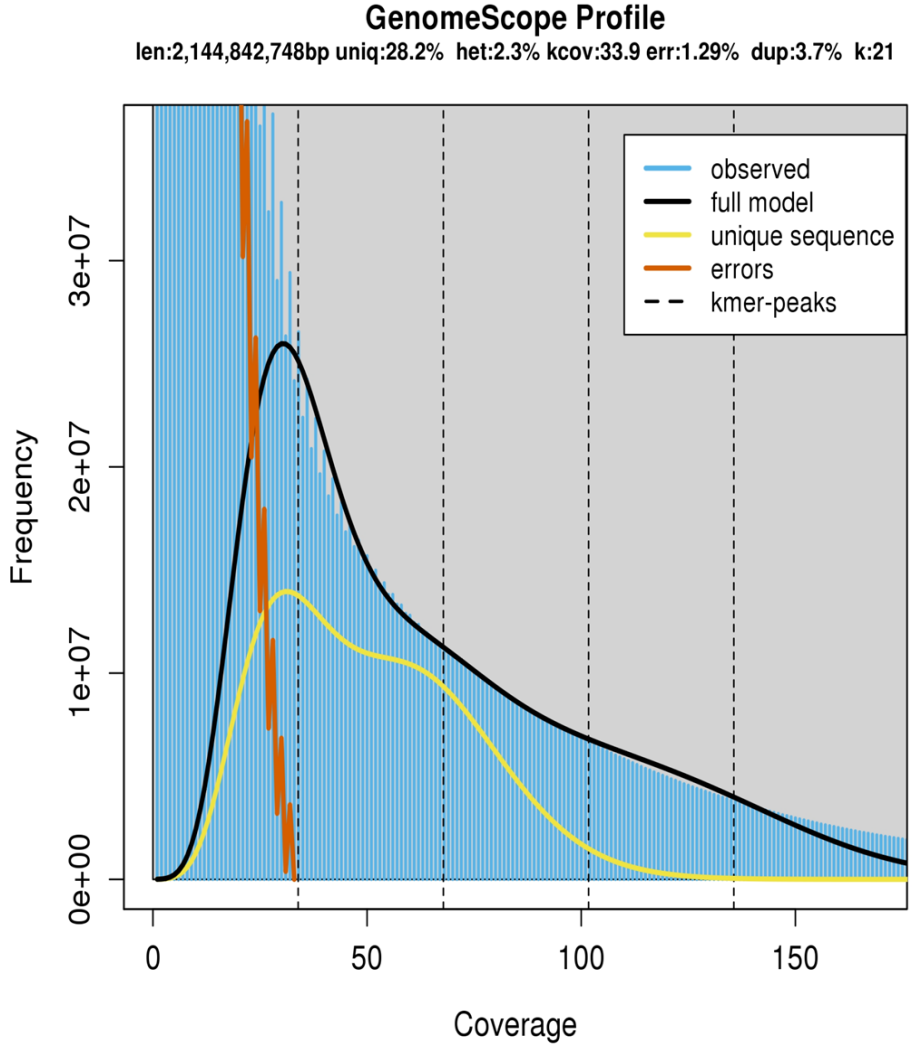
Supplementary Figure 5. Classification of genes based on their duplication type (**A**) and K_S distribution (**B**). The categories included whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), transposed duplication (TRD), and dispersed duplication (DD) genes, in C. humilis, E. guineensis and P. dactylifera.



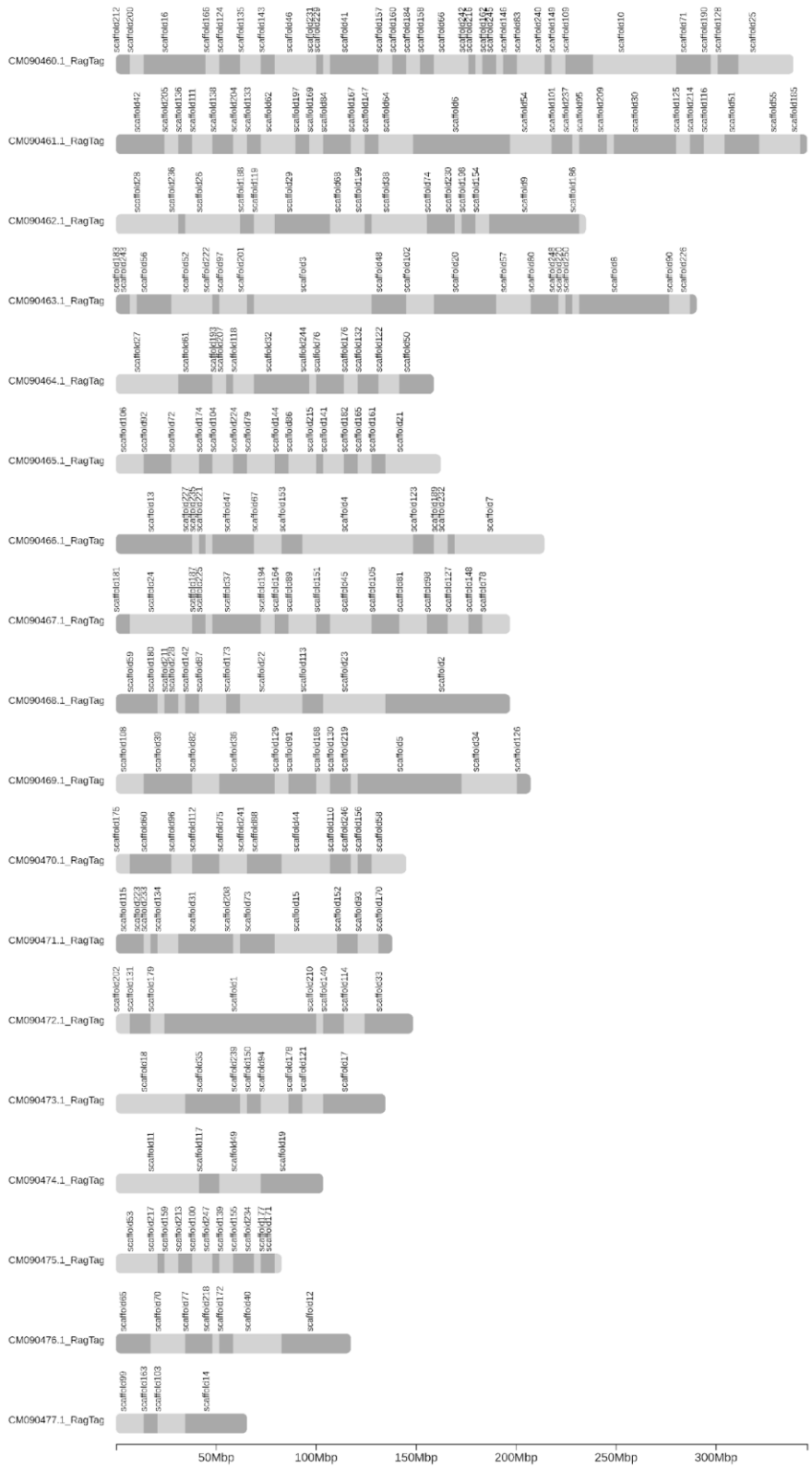
Supplementary Figure 6. Intraspecific homology collinearity levels for *P. dactylifera* **(A)** and *E. guineensis* **(B)**. The minimum length for segments is set to 100kb.



Supplementary Figure 7. The karyotype of *C. humilis* showing duplicated genes. Genes duplicated (A) seven times, (B) eight times and (C) nine times classified by duplication type: whole-genome duplication (yellow), tandem duplication (red), proximal duplication (green), transposed duplication (blue), and dispersed duplication (purple).



Supplementary Figure 8. GenomeScope analysis of *C. humilis* genome. The k-mer distribution (k-mer size 21) and the model fit are shown. The blue portion represents the analyzed k-mer frequency, while the orange and yellow lines represent errors and unique sequences, respectively



Supplementary Figure 9. Chromosome reconstruction of *C. humilis*. Each chromosome is composed of multiple scaffolds, generated previously through the integration of PacBio, Illumina, and Hi-C reads.

Supplementary Table

Supplementary Table 1. Genome annotation summary

Chromosome	Length (bp)	Nº genes	Gene bases (%)	Nº TE	TE bases (%)
Chr1	339,323,591	6,168	14.19	43,772	82.74
Chr2	345,888,753	6,215	13.48	45,215	83.45
Chr3	234,672,915	3,788	13.95	30,939	83.07
Chr4	291,554,667	5,084	13.00	37,650	83.51
Chr5	158,142,973	2,462	11.68	21,496	85.46
Chr6	163,579,036	3,247	13.33	20,765	82.18
Chr7	212,950,890	3,373	12.38	28,554	84.73
Chr8	197,431,183	3,353	12.04	26,369	83.86
Chr9	198,382,010	3,602	13.40	26,105	83.46
Chr10	206,691,416	3,538	12.78	27,415	84.07
Chr11	143,760,798	2,312	12.77	19,554	83.84
Chr12	136,708,599	2,420	12.96	17,624	83.45
Chr13	149,157,795	2,108	9.85	21,726	86.24
Chr14	133,245,200	2,114	11.09	18,141	85.15
Chr15	104,501,385	1,756	12.63	13,897	84.63
Chr16	84,500,901	1,489	12.95	11,627	84.05
Chr17	118,208,616	1,783	12.56	16,484	85.87
Chr18	65,488,747	1,181	13.71	8,670	83.29
Chr0	159,861,181	4,562	16.26	19,233	76.31
Total	**3,444,050,656**	**60,555**	**13.02**	**455,236**	**83.53**

Supplementary Table 2. Whole-genome TE annotation obtained by the EDTA pipeline

Class	Order	Superfamily	Count	Bp Masked	% masked
Class I (retrotransposons)	**LTR**	Copia	1,082,759	1,271,540,157	36.94%
		Gypsy	1,094,945	1,026,272,178	29.81%
		unknown	606,479	386,652,822	11.23%
Class II (DNA transposons)	**TIR**	CACTA	65,518	29,027,562	0.84%
		Mutator	214,013	86,753,188	2.52%
		PIF_Harbinger	19,830	5,848,576	0.17%
		Tc1_Mariner	6,389	2,428,316	0.07%
		hAT	49,197	20,887,084	0.61%
	nonTIR	Helitron	76,220	40,010,240	1.16%
Total			3,215,350	2,869,420,123	83.36%

Supplementary Table 3. Summary statistics of genome assemblies of *C. humilis*, *P. dactylifera*, *E. guineensis* and *C. nucifera*

Type	***C. humilis***	***P. dactylifera***	***E. guineensis***	***C. nucifera***
Chromosomes	18	18	16	16
Assembly size (Gb)	3.44	0.772	1.8	2.68
Longest scaffold size (Mb)	345.89	12.1	186.2	241.6
N50 (Mb)	198.38	0.897	128.31	174
Annotated genes	60,555	28,595	46,697	29,147
Repetitive DNA (Gb)	2.8	0.403	1.8	2.19
BUSCO completeness (%)	90.8	92.5	91.6	96.8
LTR Assembly Index	19.8	-	15.43	8.54
GC (%)	43.74	40.3	38.54	37.7

Supplementary Table 4. Full length LTR annotation

Super Family	Lineage	Counts
Ty1/copia	Tork	15269
	Angela	9673
	SIRE	8220
	Ale	2769
	Ivana	1548
	Ikeros	773
	Alesia	221
	TAR	117
	Gymco	38
	Alexandra	14
	Bianca	10
	Bryco	3
	Lyco	3
	Selgy	1
	Chlamyvir	1
	Ferco	1
Ty3/gypsy	CRM	11658
	Tat	9755
	Athila	4845
	Tekay	4198
	Reina	896
	Galadriel	73
	Tcn1	16
	Ferney	1

249