scientific reports

Check for updates

OPEN

Chromosome scale genome assembly and annotation of coconut cultivar Chowghat Green Dwarf

M. K. Rajesh[1,2]✉, Roli Budhwar[3], Rohit Shukla[3], Praveen Kumar Oraon[3], Shailendra Goel[4], Bobby Paul[5], Regi Jacob Thomas[6], Akshay Dinesh[1], S. Jayasekhar[1], K. P. Chandran[1], K. S. Muralikrishna[1], B. J. Nirmal Kumar[1] & Alpana Das[7]

The high-quality genome of coconut (*Cocos nucifera* L.) is a crucial resource for enhancing agronomic traits and studying genome evolution within the Arecaceae family. We sequenced the Chowghat Green Dwarf cultivar, which is resistant to the root (wilt) disease, utilizing Illumina, PacBio, ONT, and Hi-C technologies to produce a chromosome-level genome of ~2.68 Gb with a scaffold N50 of 174 Mb; approximately 97% of the genome could be anchored to 16 pseudo-molecules (2.62 Gb). In total, 34,483 protein-coding genes were annotated; the BUSCO completeness score was 96.80%, while the k-mer completeness was ~87%. The assembled genome includes 2.19 Gb (81.64%) of repetitive sequences, with long terminal repeats (LTRs) constituting the most abundant class at 53.76%. Additionally, our analysis confirms two whole-genome duplication (WGD) events in the *C. nucifera* lineage. A genome-wide analysis of LTR insertion time revealed ancient divergence and proliferation of *copia* and *gypsy* elements. In addition, 1368 RGAs were discovered in the CGD genome. We also developed a web server 'Kalpa Genome Resource' (http://210.89.54.198:3000/), to manage and store a comprehensive array of genomic data, including genome sequences, genetic markers, structural and functional annotations like metabolic pathways, and transcriptomic profiles. The web server has an embedded genome browser to analyze and visualize the genome, its genomics elements, and transcriptome data. The in-built BLAST server allows sequence homology searches against genome, annotated transcriptome & proteome sequences. The genomic dataset and the database will support comparative genome analysis and can expedite genome-driven breeding and enhancement efforts for tapping genetic gains in coconut.

Keywords Coconut genome, Transposable elements, Whole genome duplication, Genome browser, Database, Web server

Abbreviations

AED	Annotation edit distance
BUSCO	Benchmarking universal single-copy ortholog
CGD	Chowghat Green Dwarf
CLR	Complete long reads
ELMM	Exponential-log normal mixture modeling
EM	Expectation-maximization
Gb	Gigabyte
GO	Gene ontology
Hi-C	High-throughput chromosome conformation capture

[1]ICAR-Central Plantation Crops Research Institute, Kasaragod, Kerala 671124, India. [2]ICAR-Central Plantation Crops Research Institute, Regional Station, Vittal, Karnataka 574243, India. [3]Bionivid Technology [P] Limited, Bengaluru, Karnataka 560064, India. [4]Department of Botany, University of Delhi, Delhi 110007, India. [5]Department of Bioinformatics, Manipal School of Life Sciences, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India. [6]ICAR-Central Plantation Crops Research Institute, Regional Station, Kayamkulam, Kerala 690533, India. [7]ICAR-Central Plantation Crops Research Institute, Research Centre, Kahikuchi, Assam 781017, India. ✉email: rajesh.mk@icar.gov.in

HMM	Hidden Markov models
HQ	High-quality
HSP	High-scoring pair
KEGG	Kyoto Encyclopedia of Genes and Genomes
KGR	Kalpa Genome Resource
KO	KEGG functional orthologs
LTR	Long terminal repeats
NLR	Nucleotide-binding site—Leucine-rich repeat
ONT	Oxford Nanopore Technologies
RGA	Resistance gene analogues
SRA	Sequence read archives
SSR	Simple sequence repeats
TE	Transposable elements
WGD	Whole genome duplication

Coconut (*Cocos nucifera*) is a tropical monocotyledonous palm belonging to the Arecaceae family and the sole species within the genus *Cocos*. It is a vital agricultural crop, popularly known by the sobriquet '*Tree of Life*' due to its wide spectrum of uses, from nutrition to health to industrial applications. Coconuts are categorized into two main types: the 'Talls' and 'Dwarfs'. The 'Talls' are typically cross-pollinating and have a slower growth rate. The 'Dwarfs,' in contrast, are predominantly self-pollinating, grow more quickly[1], and are believed to have arisen from 'Talls' through domestication events in Southeast Asia[2]. Globally, coconut cultivation spans over 12.25 million hectares, with the Asia-Pacific region accounting for 85% of this area[3]. The Philippines, Indonesia, and India are notable for their significant contributions to coconut production and export[4]. The coconut's adaptability to coastal environments and tolerance to salinity make it a resilient crop in various tropical climates. Its economic importance is underscored by its diverse products, including coconut water, oil, fiber, and building materials.

The coconut genome is large, approximately 2.4 Gb in size[5], and rich in repeat sequences, particularly long terminal repeat retrotransposons[5,6], which play a major role in genome modification and speciation[7]. Recent efforts in coconut genomics have led to the development of a dense linkage map and the assembly of the genome into pseudomolecules, enhancing our understanding of the coconut's evolutionary history and facilitating breeding for traits such as salt tolerance and plant height[5,6].

Research has shown that the coconut genome has undergone significant transposable element invasion, possibly linked to historical sea-level fluctuations during the Pleistocene[5]. Additionally, comparative genomics has revealed divergence times between coconut subspecies and identified a conserved whole-genome duplication event in the Arecaceae family[6]. The coconut's evolutionary position as an ancient angiosperm crop makes it an intriguing subject for studying the evolution of monocot chromosomes and plant differentiation. High-quality genome assemblies of both 'Tall' and 'Dwarf' varieties[5,6,8] have been produced, providing insights into the genetic basis of trait differences, such as plant height, which is influenced by gibberellin metabolism[6]. In summary, the coconut is a multifaceted species with a complex and large genome shaped by its evolutionary history and human cultivation[9].

During the past decade, transcriptome sequencing (RNA-sequencing/ RNA-seq) has also been conducted on different coconut tissues (endosperm, leaves, zygotic embryos, and embryogenic calli) and different conditions (yellowing disease, root (wilt) disease, somatic embryogenesis)[10]. These transcriptomic resources hold immense value for the refinement of gene annotation, evaluation of gene expression patterns under different treatments, and resolution of gene regulatory networks. The genome and transcriptome sequencing initiatives in coconut have promoted the mining of sequences related to (i) genes associated with biotic/abiotic stress tolerance, e.g. resistance gene analogs (RGAs)[11], antiporter (Na^+/H^+) gene families[12], and auxin-responsive factors (ARFs)[13]; (ii) genes associated with pandan-like aroma[14], oil biosynthesis[15], flowering[16], fiber content and plant height[6]; and (iii) simple sequence repeats[17].

The genomic resources developed for coconut can shed light on its biology and adaptation and offer valuable tools for improving this essential tropical crop. We had earlier sequenced and assembled the genome of the 'Chowghat Green Dwarf' (hereafter referred to as CGD), an indigenous dwarf cultivar from India known for its resistance to root (wilt) disease (Fig. 1)[8]. The genome size of CGD was estimated to be 2.59 Gb using flow cytometry in our earlier study[8]. A hybrid genome assembly strategy was earlier employed, combining short reads from the Illumina HiSeq 4000 platform and long CLR reads from the Pacific Biosciences (PacBio) platform to overcome the challenges posed by the large genome size and high repetitive content. The draft genome assembly of CGD spanned 1.93 Gb across 26,855 scaffolds, representing approximately 75% of the estimated genome size, with an N50 of 128.74 Kb[8]. The functional annotation of predicted protein-coding genes, identification of transcription factors, transcription regulators, protein kinases, nucleotide-binding site-leucine-rich repeat (NLR) gene family, and the assembly and annotation of chloroplast and mitochondrial genomes have laid the groundwork for in-depth genetic studies and breeding programs aimed at enhancing CGD and other coconut varieties.

Here, we present an improved assembly of the CGD genome, achieved through a hybrid assembly approach that integrated the existing draft assembly, along with newly generated three datasets from long-read Oxford Nanopore Technologies (ONT) with short-read Illumina-data and Hi-C proximity ligation data-based scaffolding. The final chromosome-level assembly of 16 pseudomolecules represents a significant improvement of the CGD genome, offering a high-quality reference for the scientific community. This assembly has been meticulously validated and annotated, revealing a rich tapestry of genes, including those involved in gibberellic acid and oil biosynthesis pathways and non-coding RNA elements contributing to the genome's regulatory

Fig. 1. Typical palm of Chowghat Green Dwarf (CGD) coconut cultivar.

complexity. Through synteny analysis with related palm species and a detailed examination of gene duplication events, we provide insights into the evolutionary history and genomic architecture of the coconut palm. The genomic resource presented herein provides a valuable resource for the coconut research community. It lays the groundwork for targeted genetic improvement by exploring the genetics underlying key agronomic traits, enables variant discovery, and can speed up breeding programs by unlocking the full agronomic potential of this vital tropical crop.

The availability of these genome and transcriptome datasets necessitates the development of an integrated database that can be used to store, analyze, and mine these datasets for coconut researchers, especially breeders. We have built the dynamic 'Kalpa Genome Resource' web server, which integrates the coconut genome of the CGD cultivar, transcriptome profiles, and annotation datasets. In addition, we have predicted gene models and functionally annotated them. Also, we have used the available RNA-Seq datasets and included gene expression profiles of these datasets. The 'Kalpa Genome Resource' was developed using technologies like NodeJS, Express JS, MongoDB, SQLite, and JavaScript. This genome browser, supported by user-friendly modules and interfaces, will enable easy retrieval of gene families and their associated annotations utilizing a repertoire of databases, viz., Interpro, KEGG, and GO (Gene Ontology). Tools such as BLAST and JBrowse (genome browser) have been included for easy visualization of features of the CGD genome and its comparative analysis. The genome browser will be updated regularly with novel genome sequences and annotations.

Materials and methods

The schematic workflow of CGD genome assembly and annotation is provided in Supplementary Fig. 1.

Sampling and genome sequencing

High-quality genomic DNA was extracted from spindle leaves of a representative CGD palm (Accession IND029; IC296656) cultivated at the ICAR-CPCRI Experimental Farm, Kasaragod District, Kerala State, India, employing the Qiagen MagAttract HMW DNA Kit (Qiagen, Germany). After isolation, genomic DNA underwent a 2.0× bead purification process to enhance sample purity and was quantified utilizing the dsDNA Broad Range (BR) Qubit assay (Thermo Fisher Scientific, USA).

Sequencing was carried out in Illumina and Oxford Nanopore Technologies (ONT) sequencing platforms. The extracted DNA underwent library preparation using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, USA), followed by short-read Illumina whole-genome sequencing (WGS) at a coverage depth of ~ 33X. The sequencing was carried out using the Illumina NovaSeq 6000 platform, employing Illumina TruSeq PCR-free chemistry.

For ONT-based long-read sequencing, 1 µg of high molecular weight genomic DNA underwent library preparation using the DNA V14 kit (SQK-LSK-114; Oxford Nanopore Technology, UK), following the manufacturer's recommendation. Sequencing was performed using three flow cells on the MinION™ Mk1C sequencing platform.

The Hi-C library construction was conducted using the Arima-HiC kit v1 (A510008, Arima Genomics, USA) per the manufacturer's protocols. Initially, crosslinked DNA underwent digestion with a mixture of Arima restriction enzymes. The resulting fragments with 5′ overhangs were then combined with biotinylated nucleotides, followed by ligation of the blunt ends. Subsequently, DNA was purified, sheared, and enriched using streptavidin beads. Sequencing libraries were prepared using compatible Illumina adapters with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, USA). These libraries underwent paired-end sequencing on the Illumina NovaSeq 6000 platform, resulting in a total output of ~ 123 Gb (47X coverage) of Hi-C data.

Genome assembly and scaffolding

The primary assembly of the CGD genome was constructed using a hybrid strategy that combined long-read sequencing data from the Oxford Nanopore Technologies (ONT) platform and short-read sequencing data from the Illumina platform. High-quality (HQ) Illumina reads were selected after quality control using the fastp tool[18] with a cutoff Phred score of 30 and a minimum read length of 70 nucleotides. The contig level assembly was constructed using MaSuRCA(v 4.0.1)[19], incorporating both filtered Nanopore and Illumina reads (current study) along with PacBio data generated in the previous study (SRA Accession No. SRS2696501).

The resulting contig assembly underwent filtering to remove haplotigs, a process accomplished with purge_haplotigs (v1.1.1)[20], and scaffolds shorter than 1000 bp were discarded. This assembly yielded 19,612 contigs, with a cumulative length of approximately 1.9 Gb. To enhance the contiguity of the assembly, we employed reference-based scaffolding, carried out using RagTag (v2.1.0)[21], with the genome assembly of the Chinese Dwarf cultivar as a reference (National Genomics Data Center, Beijing Institute of Genomics, Accession no. GWHBEBU00000000)[6]. The primary assembly was then scaffolded with Hi-C data using the YaHS tool (v 1.1)[22]; the Hi-C data quality control was carried out using HiCUP (v 0.7.4)[23].

Assessment of genome assembly

The completeness of the genome assembly was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool (v5.4.7)[24] using the *Embryophyta* lineage. High-quality whole-genome sequencing (WGS) Illumina short reads were aligned to the genome using Bowtie2 (v2.4.1)[25] to determine the mapping coverage onto the assembled genome. A *k-mer*-based assessment of completeness and per-base accuracy [Quality Value (QV)] score was performed using Merqury (v 1.3)[26] to estimate the base-level accuracy of the assembled genome. The long-terminal repeat (LTR) assembly index (LAI) was also calculated by LTR_retriever (v.2.9.0)[27] following default parameters to assess the assembly quality.

Identification of telomeres

Telomeric regions within the CGD genome were identified using the Telomere Identification Toolkit [Tidk (v0.2.0); https://github.com/tolkit/telomeric-identifier]. We employed the -c Poales option to focus on monocot-specific repeats.

Repeat identification and annotation

We employed RepeatModeler (v2.0.4)[28] and LTR_retriever (v2.9.0) to de novo predict Transposable Elements (TEs) in the assembled coconut genome. Repbase Database (v28.06)[29] (https://www.girinst.org/repbase/) was used to annotate the identified TEs. The resulting TE library was then used with RepeatMasker (v4.1.5)[30] to identify and soft mask the TEs within the draft assembly.

The genome underwent microsatellite screening using the MIcroSAtellite v2.1 (http://pgrc.ipk-gatersleben.de/misa/) to detect putative perfect mono- to hexa-nucleotide repeats as well as complex microsatellites under default parameters. Microsatellite detection was performed using predefined parameters: unit sizes ranged from 1 to 6 nucleotides, with repeat thresholds set as follows: mononucleotides required 1 to 10 repeats, dinucleotides 2 to 9 repeats, trinucleotides 3 to 6 repeats, and tetranucleotides, pentanucleotides, and hexanucleotides required 4 to 5, 5 to 5, and 6 to 5 repeats, respectively. The search protocol also permitted interruptions of up to 100 base pairs between adjacent microsatellites. The masked assembly served as the basis for further gene prediction steps.

A genome-wide estimation of the insertion times for *copia*, *gypsy*, and unknown intact LTR elements in the CGD genome was conducted using Extensive De novo Transposable element Annotator (EDTA)[31], which identified intact LTRs across the genome using –sensitive 1 –anno 1 parameters. Along with this, the insertion time (T) of each LTR sequence was calculated using the formula $T = K/2\mu$, or $T = (1 - \text{identity}) / 2\mu$, applying a substitution rate (μ) of 1.3×10^{-8}, specified via '-μ' parameter within the EDTA tool.

Gene prediction and functional annotation

Gene prediction was carried out using the MAKER2 pipeline (v3.01.03)[32], which utilizes tools such as GeneMark EP (v4.71)[33] and Augustus (v3.3.2)[34]. We combined de novo predictions with homology information from RNA-Seq data to refine our gene predictions. Transcript sequences from in-house generated RNA-Seq data of embryogenic calli [NCBI Short Read Archive (SRA) Accession no. SRX472157][35], endosperm calli (SRX10625025)[36], and leaf tissues (SRX436961 and SRX437650)[37] were assembled de novo and merged to create a non-redundant set of transcript sequences as EST evidence. Protein sequences from the oil palm (Refseq database accession GCF_000442705.1) were also used as evidence to obtain the final gene models. The obtained gene models were filtered to retain only those genes with an AED (Annotation Edit Distance) score of less than 0.5 and a minimum protein product length of 100 amino acids.

Functional annotation of the gene models was performed by performing a BLASTp comparison against a database of non-redundant peptide sequences from the Arecaceae family obtained from the NCBI RefSeq database[38]. The BLAST results were filtered to retain the best-hit HSP (High-Scoring Pair). Interproscan (v 5.64-96.0)[39] was used to annotate protein domains in the gene products of predicted genes. Gene Ontology (GO) terms were assigned with PANNZER2 web server[40]. Protein sequences of the predicted genes were used as an input in GhostKOALA (KEGG mapper) to assign KEGG functional orthologs (KO) to the predicted genes[41]. The KEGG ortholog IDs were used as input in the KEGG reconstruct tool to obtain pathway IDs.

Identification of non-coding RNAs

We employed several tools to predict non-coding RNAs from the assembled genome assembly. To identify tRNAs within the genome, we utilized tRNAscan-SE (v2.0.9)[42], setting a cutoff score of 50 to obtain the final predictions of tRNAs in the genome. The prediction of rRNAs was accomplished using Barrnap (v0.9) (https://github.com/tseemann/barrnap/), with the eukaryotic option enabled. We conducted an Infernal search[43] against Rfam database models[44], using default parameters to identify other non-coding RNA species. Only complete (not fragmented) predictions with an E-value of less than 0.05 were retained as non-coding RNAs.

Prediction of disease resistance R genes

The Nucleotide-Binding Site—Leucine-Rich Repeat (NBS-LRR) gene family is generally classified based on the presence of two crucial domains: NB-ARC (IPR002182) and LRR (IPR032675). To identify these domains within the proteins of predicted CGD gene models, we conducted sequence pattern mining using HMMscan (https://www.ebi.ac.uk/jdispatcher/pfa/hmmer3_hmmscan) utilizing hidden Markov Models from the Pfam database. We also incorporated predictions from the NLR annotator (https://github.com/steuernb/NLR-Annotator) to compile a list of potential NLR genes. Additionally, CGD genome resistance (*R*) genes were analyzed using the DRAGO-apitool (https://github.com/sequentiabiotech/DRAGO2-API) against PRGdb v4.0[45]. This approach enabled the identification of the potential repertoire of resistant gene analogs (RGAs), which were subsequently classified into various categories according to their specific characteristic domains, viz., coiled-coil (CC) (CC)–nucleotide-binding site (NBS)–leucine-rich repeat (LRR) (CNL), Kinase (KIN), serine/threonine–LRR (receptor-like protein [RLP]), receptor-like kinase (RLK), and other categories. A comparison was also made with the RGAs predicted from the Catigan Green Dwarf genome (Accession no. QRFJ01000000).

In addition to identifying NLR genes within predicted gene loci, we sought to locate genomic loci with NLR domains beyond the predicted genes. We scanned the loci of the NLR genes obtained previously using the MEME tool from MEME-Suite to predict motifs within these sequences. We then employed the MAST tool from MEME-Suite to search the genome for similar loci using the predicted motifs. Sequences from the identified loci were translated in all six frames using Transeq from the EMBOSS toolkit and subsequently analyzed using InterProScan. A non-redundant list of NLR loci was compiled by combining the NLR loci predicted as genes with those obtained from MAST.

Multiple sequence alignment was conducted using MAFFT[46] with default parameters to investigate the evolutionary relationships among RGA family members. The alignment file was then taken forward to construct a phylogenetic tree using IQ-TREE (v2.3.6)[47], employing the in-built model finder and maximum likelihood

(ML) method with 1000 bootstrap replicates for support. The resulting tree was finally visualized using the web-based ITOL[48] tool.

Prediction of genes involved in gibberellic acid and oil biosynthesis

We have identified genesassociated with oil and gibberellic acid (GA) biosynthesis pathways in the CGD genome. Firstly, we retrieved all the sequences based on keywords related to these two biosynthetic pathways from the NCBI, and the organism search was restricted only to Viridiplantae (NCBI: txid 33090). The downloaded peptide sequences served as a query against CGD-predicted protein sequences. Both BLASTp and BLASTn homology-based searches were performed with an E-value cutoff of 1×10^{-5}. Secondly, we manually curated the BLAST results and functional annotation file to finalize the final list of genes related to oil and GA biosynthetic pathways.

Whole genome duplication (WGD) analysis

To determine whether the whole genome duplications (WGDs) events identified in other coconut genomes correspond to the same or different evolutionary events as those reported in related plant genomes[6], we investigated the order of these WGDs. We achieved this by extracting co-linear paralogous pairs from the WGDs within the genomes of coconut and three representative palms, viz., *Elaeis guineensis*, *Areca catechu*, and *Calamus simplicifolius*. We employed the wgd2 pipeline to extract paralogous gene pairs within each genome using "wgd-dmd"[49]. Subsequently, Ks values were calculated for each paralogous gene pair using "wgd-ksd." After that, individual results were analyzed (for each genome one by one). The exponential-lognormal mixture modeling (ELMM) approach was then utilized within the "wgd-viz" function to detect the Ks peaks in default parameters with a maximum number of EM (Expectation-Maximization) iterations and a maximum number of EM initializations to determine the acceptable peaks and relative height at which the peak width is measured. The "wgd peak" function was used to identify and place the WGD event. The bootstrap value was set at 200, with default parameters involving the confidence interval ('ci' method). However, to validate the peak sensitivity, we applied two other methods within the "wgd peak" function, i.e., the highest density region (HDR) method using the "hdr" argument and the heuristic method by specifying "heuristic" argument in the command. All these methods supported our peaks with no deviations, confirming the robustness of our peak detection. The kernel density distribution curve of K_s was created using the ggplot2 library in R with the kernel smoothing density function (geom_density) to visualize all the peaks from each genome.

Comparative genomics

Syntenic relationships at the gene level between the current coconut genome assembly and the genomes of oil palm and date palm were investigated using the McScanX tool[50]. Conserved syntenic blocks between the coconut, date palm (GCF_009389715.1), and oil palm (GCF_000442705.1) genomes were determined to identify similarities between chromosomes. Briefly, BLASTp search of protein sequences of predicted gene models with oil palm and date palm peptide sequences followed by McScanX (-g 3 -m 15 -s 10) algorithm determined the syntenic blocks between the coconut genome and oil palm and date palm genomes. Additionally, whole genome alignment was performed using the Dgenies tool[51] between the CGD genome (current study) and published Chinese Tall (CNGB accession-GWHBEBT00000000) and Chinese Dwarf (CNGB accession-GWHBEBU00000000) coconut genome assemblies[6].

Gene expression profiles

We have used in-house generated coconut transcriptome data, which include data generated by RNA-Seq of leaves of healthy and root (wilt) diseased CGD palms [NCBI Short Read Archive (SRA) accession number's SRX436961 and SRX437650[37], respectively], embryogenic calli of West Coast Tall cultivar (SRX472157)[35] and leaves of Chowghat Orange Dwarf (COD) cultivar in response to infection by *Phytophthora palmivora* at different time-points (SRX591407 to SRX591412; SRX9197106 to SRX9197111)[52]. The raw files were mapped to the assembled genome using HiSAT2[53], and transcript models were built from the three datasets using StringTie[54]. SeqPing predicted gene annotation file was used as a reference for creating the transcript models. The predicted proteins from SeqPing were used as a query to perform tBLASTn[55] against transcript models. Proteins hitting a BLAST, with query coverage and percentage identity of $>=90\%$, were regarded to possess protein-coding evidence in the transcript models.

Database implementation

The 'Kalpa Genome Resource' (KGR) comprises a web-based front-end interface and a back-end application server. To enhance the user experience and to improve the website interface, the core of KGR was developed by using a suite of five basic core modules or tech stacks, viz., (i) Node.js (https://nodejs.org/en); (ii) Express JS (https://expressjs.com/); (iii) MongoDB (https://www.mongodb.com/); (iv) SQLite (https://www.sqlite.org/), and (v) JavaScripts, for processing, analysis and visualization of data. Mongo DB was used to hold the core information of the database, SQLite was used to integrate computer-intensive database querying applications and HTML and Java scripts were used to maintain the basic database framework. Additionally, the KGR incorporates JBrowse (https://jbrowse.org/jb2/), a comprehensive genome browser that is open-source and built using JavaScript and HTML5, and a BLAST online search tool. We have introduced user log-in features with Google, Microsoft, Github, Twitter, e-mail, and SMS services.

Results and discussion

Genome sequencing

To achieve chromosomal level assembly of Chowghat Green Dwarf (CGD) cultivar, four sequencing technologies were integrated: ONT long reads (Oxford Nanopore), paired-end short-read Illumina sequencing, and Hi-C proximity ligation data, along with the PacBio long read data generated in our previous study[9]. The Illumina library yielded 86.46 Gb of raw data with 82.40% high-quality (HQ) reads after quality control (Supplementary Table 1). The ONT library produced 23.80 Gb of raw data with an N50 read length of 16,245 bp (Supplementary Table 2). The Hi-C library yielded 123.09 Gb of data with 97.41% high-quality reads (Supplementary Table 3).

Genome assembly and scaffolding

We performed a hybrid assembly approach to generate the contig-level genome of CGD by integrating raw data generated through ONT, PacBio, and Illumina sequencers. Using the MaSuRCA assembly pipeline, we first crafted 23.80 Gb, error-corrected ONT long reads. These reads boasted an average length of 4,255 bp and an N50 of 16.2 Kb. We then incorporated error-corrected PacBio long reads and combined them with ONT long read data. Additionally, we integrated ~ 471.8 million Illumina short reads to refine the accuracy of the assembly further. This multi-layered strategy culminated in the primary 'CGD' genome assembly, encompassing ~ 2.68 Gb. This primary assembly was structured into 44,376 contigs, with the N50 standing at 122 Kb and N90 at 34 Kb.

We then employed a reference-based assembly approach using an already published high-quality coconut dwarf genome (Accession no. GWHBEBU00000000)[6] to lay the groundwork for our genome assembly. Following this, we integrated Hi-C sequencing data, which allowed us to anchor ~ 2.62 Gb (~ 97%) of the genome to 16 pseudomolecules with an additional ~ 65 Mb unplaced genome data (1914 scaffolds) (Fig. 2A, B; Table 1). Scaffolds with less than 1000 bp length were removed. These pseudo-molecules were named based on their sizes. The chromosomes varied in length, with the longest being ~ 241 Mb and the shortest ~ 90 Mb, achieving an N50 value of ~ 174 Mb (Supplementary Table 4).

The current version of the genome assembly of the CGD cultivar has significantly improved the earlier published CGD genome[9], which was assembled using short-read Illumina and long-read PacBio data. The current high-quality assembly comprises 16 pseudomolecules, representing ~ 97% of the CGD genome. From

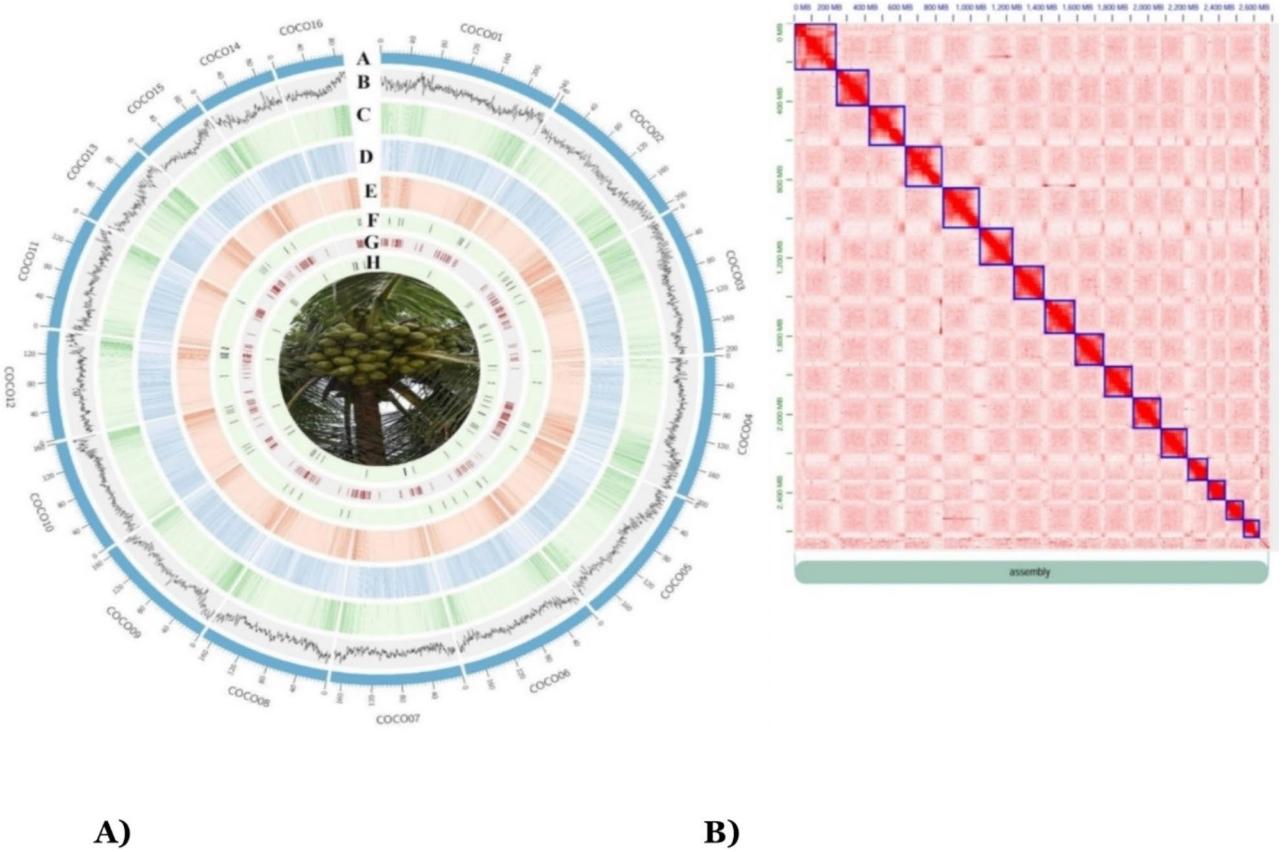


Fig. 2. Assembly of Chowghat Green Dwarf genome indicating different genomic features. (**A**) The circles from outside to inside represent: (A) physical map; (B) z-score normalized GC content; (C) gene density; (D) TE density; (E) SSR density; (F) Gibberellic acid biosynthesis genes; (G) oil biosynthesis genes; and (H) NLR loci. The pseudochromosomes are plotted in a unit of 1 Mb. (**B**) Hi-C image at 5 Mb resolution of the final CGD Genome after manual curation.

	CGD version 1.0 (Rajesh et al., 2020)	CGD version 2.0 (current study)
Sequencing technologies	SMRT PacBio RSII and Illumina sequencing	Illumina, ONT, SMRT PacBio RS II, and HiC
Total number of scaffolds	26,885	16 pseudomolecules (1912 unplaced scaffolds)
Assembly size (Gb)	~1.93	~2.68
Longest scaffold size (Mb)	~1.26	~241.6
N50 (Mb)	0.128	~174
BUSCO completeness (%)	84.6	96.80
Quality value (QV)	-	32.30
LAI index	-	8.54
Predicted gene models	13,707	34,483
Annotated genes	5980	29,147
tRNA genes	470	1676
rRNA genes	463	1383
Repetitive DNA (Gb)	~1.48	~2.19
GC (%)	37.59	37.70

Table 1. A comparative account of the genome of coconut cultivar Chowghat Green Dwarf.

the 26,855 scaffolds with an N50 value of ~128 Kb and final genome size of ~1.93 Gb reported in our earlier study[9], the current genome assembly consists of 16 pseudomolecules equating to a final assembly size of ~2.68 Gb (including unplaced scaffolds) with a N50 value of ~174 Mb.

Assessment of genome assembly

To assess the quality and completeness of our genome assembly, we employed BUSCO analysis using the *embryophyta_odb10* database. As presented in Supplementary Fig. 2, the CGD assembly scored 96.80% and 1.80% for complete and fragmented BUSCO assessments, respectively. Genome completeness using BUSCO improved from ~84% (previous assembly; Rajesh et al., 2020) to ~96.80% (present assembly).

Additionally, we aligned the clean genomic reads from Illumina libraries to our assembled CGD genome. Approximately 93.7% of the short Illumina reads could be aligned to the CGD assembly. These results highlight the solidity and integrity of our genome assembly, affirming its aptness for subsequent genome characterization and annotation. The contiguity, completeness, and accuracy of the genome were evaluated by *k-mer* completeness score of 87.35% (Supplementary Fig. 3), consensus per-base accuracy (QV; 32.30) (Supplementary Table 6), and LTR assembly index of 8.85 (Table 1). These indices highlight the enhanced quality of the current CGD genome assembly.

Identification of telomeres

Telomeric repeat sequences were identified at both ends of the CGD genome in five chromosomes; nine chromosomes showed telomere signals at one end, and two chromosomes did not show any signatures of telomeric presence, which suggests the potential for further improvement (Supplementary Fig. 4).

Repeat identification and annotation

We delved into characterizing repetitive elements in the CGD genome using a blend of de novo and homology-based approaches using the Repbase database. Repetitive DNA sequences accounted for about 2.19 Gb (81.64%) of the CGD genome (Supplementary Table 5), which is reflected in the enlarged genome size of coconut in comparison to date palm (671 Mb)[56] and oil palm (1.8 Gb)[57] genomes. Among the repetitive elements, LTR elements were the most prevalent, constituting 53.76% of the genome. Within the LTR class, LTR *copia* and LTR *gypsy* were the most abundant elements, comprising 40.57% and 12.98% of the genome, respectively. These observations corroborate the reports of an abundance of *copia* and *gypsy* elements in palm genomes[58]. Also, the abundance of these repetitive elements, particularly LTR elements, underscores their significant influence on shaping the genomic landscape of coconut palms (Fig. 3).

The draft genome assembly screening identified 592,389 putative microsatellite loci across 1,928 sequences, with 433,941 mononucleotide repeats. Among the remaining simple sequence repeats (SSRs) loci, dinucleotide repeats were the most prevalent, with 92,555 instances, accounting for approximately 15.6% of the total. Additionally, 90,238 complex simple sequence repeats were discovered (Results available in Figshare).

EDTA analysis identified 9142 intact LTR retrotransposons within the CGD genome. Genome-wide LTR insertion time analysis from these intact LTRs showed proliferation of *copia* and *gypsy* elements in the CGD genome occurred during distinct periods. The *copia* elements exhibited recent activity, most proliferating within the last 2 million years. In contrast, *gypsy* elements showed evidence of ancient activity, with divergence occurring between 2 and 6 million years ago (Fig. 4A); these results are consistent with previous studies[59].

Gene prediction and functional annotation

The gene prediction pipeline, MAKER, identified 34,483 protein-coding genes in the CGD genome, with an average gene length of 6,541 bp and 5.9 exons per mRNA (shown in Figshare). Through homology searches

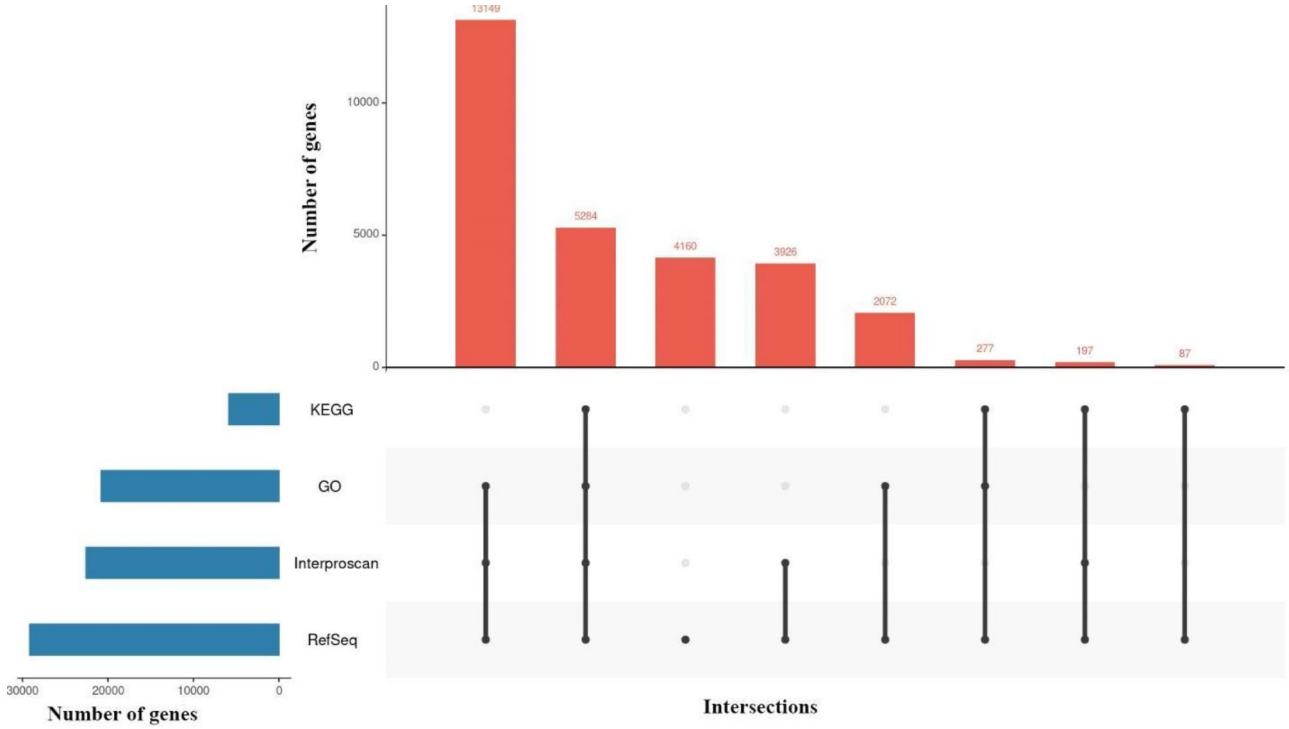


Fig. 3. Functional annotation of the CGD v2 protein-coding genes. The upset plot details the unique and overlapping annotations contributed by RefSeq, Interproscan, GO, and KEGG.

against the RefSeq database, we successfully annotated 29,147 genes, representing approximately 84.5% of the predicted genes (Fig. 3). Meanwhile, InterProScan provided domain annotations for 22,556 CGD genes (data available in Figshare).

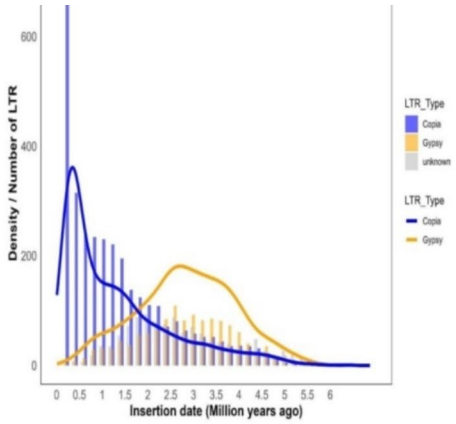
Identification of non-coding RNAs

Our examination of the CGD genome identified a diverse range of non-coding RNAs (ncRNAs). Specifically, we detected 1,676 tRNAs and 65 copies of 18S rRNA, 70 copies of 28S rRNA, 74 copies of 5.8S rRNA, and 1,174 copies of 5S rRNA. Additionally, a substantial number of other ncRNAs were found, totaling 20,433 (data available in Figshare). This comprehensive identification offers a detailed view of the ncRNA landscape within the coconut genome, underscoring its complexity and functional diversity.

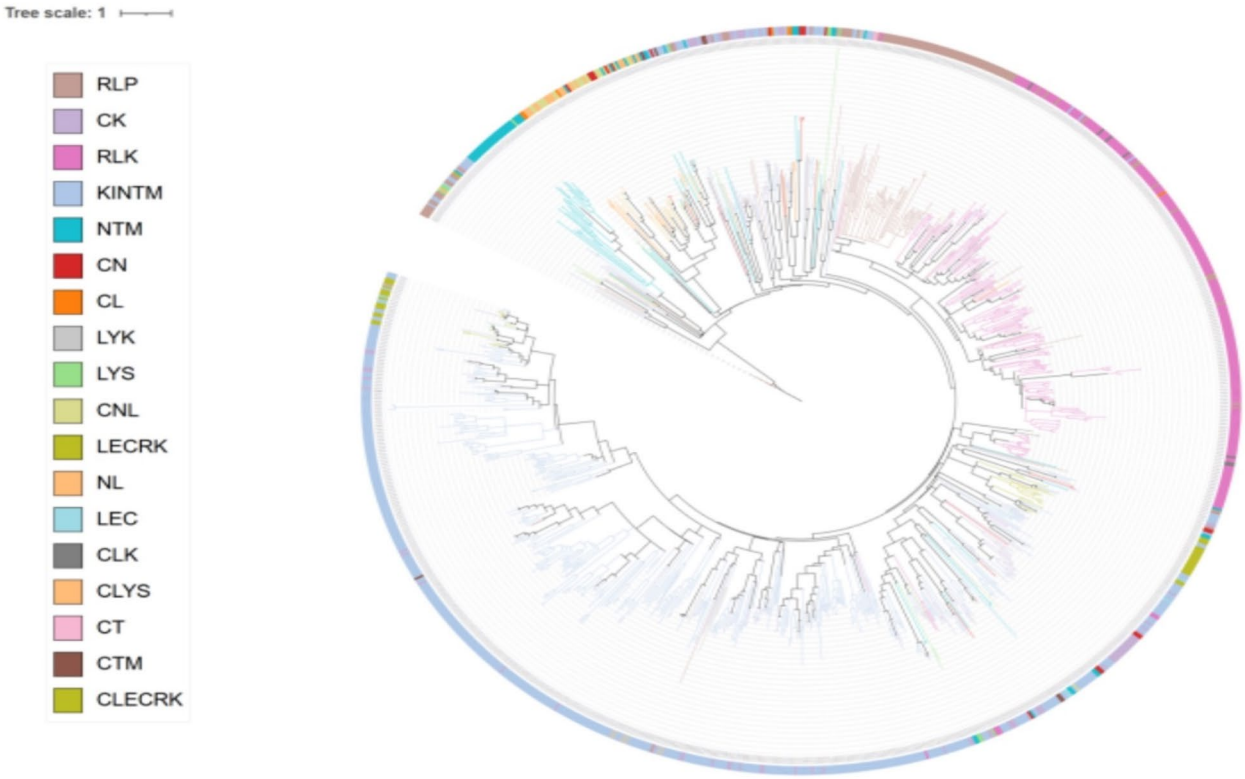
Prediction of resistance gene analogs (RGAs)

To identify resistance genes in the CGD genome, we conducted an in-depth genome-wide prediction of RGAs, resulting in the identification of 1368 RGAs. The predicted RGAs were grouped into 24 primary categories, determined by their internal domain structures and motif arrangements (Table 2). A survey of RGAs revealed that the KIN domain-containing gene family dominated the CGD genome; there were 133 single kinase domains and 505 kinase-transmembrane domain (KIN-TM) containing gene families. Around 237 members represented the RLK gene family, while the RLP family had 112 members. Other categories include NBS- transmembrane (N-TM) with a count of 58, lectin-receptor kinase (LecRK) with 35, and cytoplasmic kinase (CK) with 83 genes. Minor categories such as coiled-coil lysM domain receptor (CLys), Coiled Coil-Toll/interleukin-1 receptor (CT), and coiled-coil lectin-receptor kinase (CLecRK) had comparatively low count, indicating their limited presence in the coconut genome. In addition, CN and CNL genes were represented by 20 and 22 genes, respectively, while NL genes had 23 occurrences.

This distribution highlights the diversity and relative abundance of different *R* gene classes in the CGD genome, with KIN and RLK-related genes playing a predominant role. Phylogenetic analysis of resistance gene analogs (RGAs) in the CGD genome revealed three major clades based on domain structures and evolutionary relationships[11,59]. The first clade consists of NBS-containing genes, including Coiled Coil- Nucleotide-binding site (CN), Coiled Coil- Nucleotide-binding site- Leucine-rich repeat (CNL), Nucleotide-binding site- Leucine-rich repeat (NL), Nucleotide-binding site (N), and Nucleotide-binding site-transmembrane (NTM), which clustered together, indicating their shared role in pathogen recognition and defense mechanisms. The second clade encompasses the kinase-related families, including KIN, KINTM, RLK, and RLP, highlighting their collaborative functions in signal transduction and stress responses. The third clade includes other R gene classes with lower counts, such as C, CK, CL, CLECRK, CLK, CLYS, CT, CTM, L, LEC, LECRK, LYK, LYS, T, and TRAN. These findings reveal the diversity and functional specialization of RGAs, contributing to the coconut palm's resilience against biotic challenges. Compared to the Catigan Green Dwarf genome[59], a significantly higher number of RGAs were identified in the present study (Table 2). Our finding indicates three major clades



A)



B)

Fig. 4. Genome-wide analysis within CGD genome of intact LTRs and RGAs analysis. (**A**) Insertion Time analysis of intact LTRs within the CGD genome. (**B**) A phylogenetic tree was generated using the maximum likelihood method in IQ-TREE, derived from the sequence alignment of all predicted RGAs found in the CGD genome assembly.

from phylogenetic analysis NBS containing genes, RLK and RLP gene family in one clade together and KIN-TM as a major third clade all the other categories fallen within/between these three major clades (Fig. 4B). The availability of whole genome sequences of diverse coconut accessions will enable a deeper understanding of the evolution of RGAs and support breeding for disease-resistant varieties.

Prediction of genes involved in gibberellic acid (GA) and oil biosynthesis

Gibberellins represent a critical class of phytohormones that play a pivotal role in enhancing plant growth, developmental processes, and extension of post-harvest longevity in horticultural crops[60]. These hormones

Class	CGD genome count	CATD genome count
C	4	-
CK	83	-
CL	7	-
CLECRK	1	-
CLK	8	-
CLYS	1	-
CN	20	16
CNL	22	90
CT	2	-
CTM	6	192
KIN	133	-
KINTM	505	-
L	16	-
LEC	7	-
LECRK	35	-
LYK	12	-
LYS	20	-
N	7	-
NL	23	34
NTM	58	-
RLK	237	-
RLP	112	-
T	1	2
TRAN	48	-
Total	1368	334

Table 2. NBS-LRR disease resistance genes in the genomes of Chowghat Green Dwarf (CGD) and Catigan Green Dwarf (CATD)[59].

contribute to increasing the resilience of crops to a repertoire of stresses by altering the expression of genes associated with antioxidant systems (both enzymatic and non-enzymatic), osmoprotectants, as well as various proteins and enzymes[60]. In addition, gibberellins function synergistically with other plant growth regulators, thereby facilitating augmented growth dynamics and physiological activities in plants[60,61]. Seventy-seven genes related to GA biosynthetic pathways (Fig. 2 layer F) were identified in the CGD genome (data available in Figshare).

Oil biosynthesis is a multifaceted physio-biochemical process. In the past few years, significant advances have been made in elucidating the biochemical pathways that mediate oil synthesis. As a result of these developments, numerous pivotal enzyme genes implicated in oil biosynthesis have been successfully isolated and characterized; these achievements have paved the way for modest advancements in leveraging gene function technologies to augment oil content within seeds or to refine the profile of fatty acids[62]. Our understanding remains limited regarding transcriptional regulatory factors influencing genes associated with lipid biosynthesis in coconut. Furthermore, knowledge of how key enzymes might synergistically coordinate and regulate synthetic and metabolic routes is still being determined. Investigating these elusive mechanisms promises to reveal further regulatory insights and strategies to enhance lipid accumulation in coconut. In total, 444 genes related to oil biosynthetic pathways were mined from the current CGD genome (Fig. 2 layer G). Further studies on oil metabolism gene expression profiles in coconut might shed light on the co-expression networks pertaining to acyl metabolism-related genes.

Whole genome duplication (WGD) analysis

WGD events were analyzed by calculating the Ks values between paralogous gene pairs within C. *nucifera*, *E. guineensis*, *A. catechu*, and *C. simplicifolius*. The Ks distribution analysis revealed that all compared Arecaceae members analyzed experienced two rounds of WGDs, τ and ω (Fig. 5; Supplementary Fig. 5). Evolutionary dating based on Ks values indicated similar occurrence times as previously described[6]. Two rounds of WGD events have occurred in all palms since the divergence of monocots and eudicots[56,63]. The first event, τ-WGD, is shared by nearly all monocots, except Alismatales and Acorales, and occurred at ~150 Mya[63]. The second event, ω-WGD, unique to palm species, occurred at ~75 Mya, resulting in palaeotetraploidy. In the current study, we observed two distinct peaks in the Ks distribution. This suggests that *C. nucifera* has experienced two whole genome duplication events during its evolution, an observation reported earlier in the Chinese coconut genome[5].

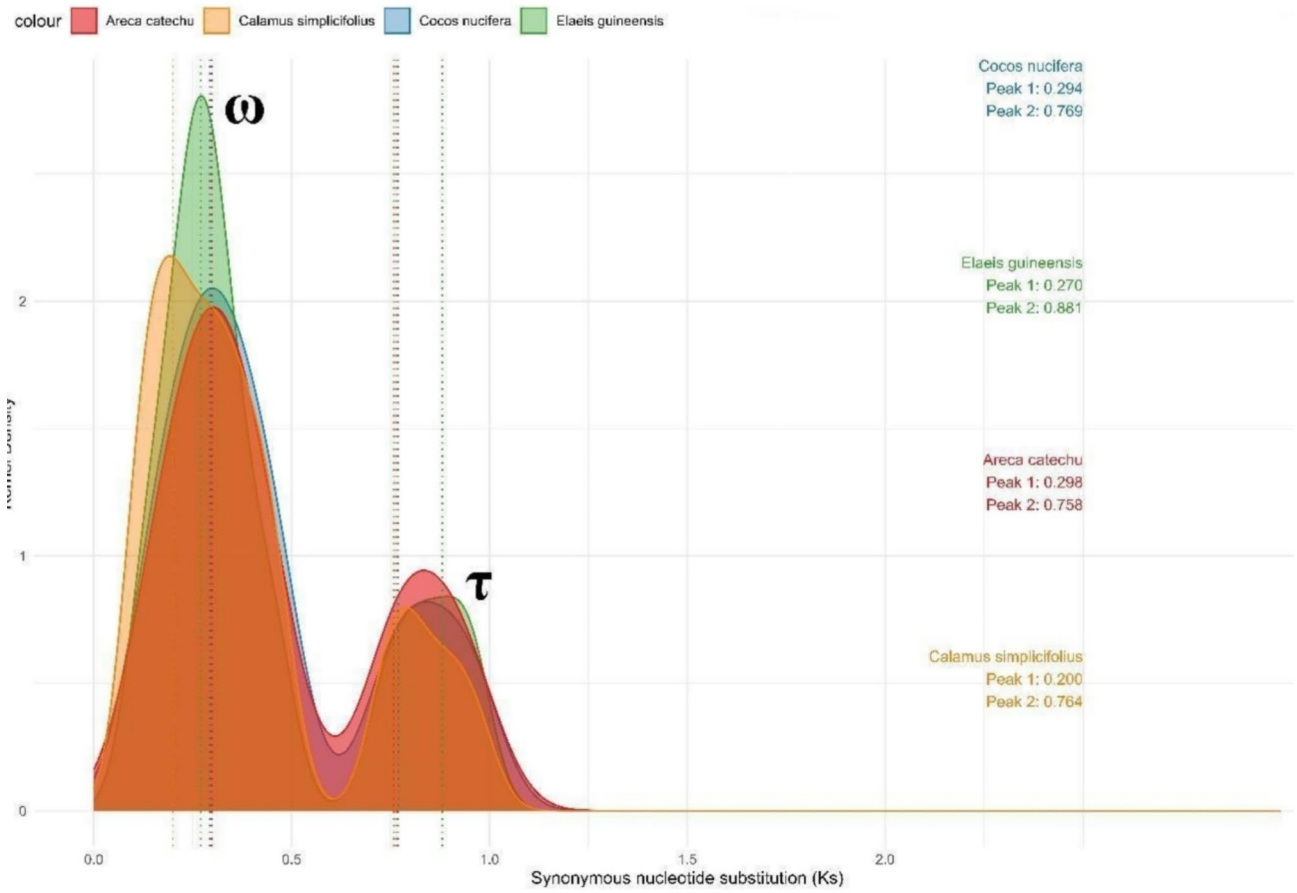


Fig. 5. The density distribution of synonymous nucleotide substitutions (Ks) in whole genome duplication analysis.

Comparative genomics

The gene synteny analysis results revealed that the CGD genome exhibited significant gene structure and order conservation, sharing 447 syntenic blocks with the oil palm genome and 489 syntenic blocks with the date palm genome (Supplementary Fig. 6). This provides insights into the evolutionary relationships and genomic synteny among these palm species.

Whole-genome alignment using the Dgenies tool revealed a one-to-one syntenic relationship between the CGD genome from this study and the pre.

viously published Chinese Tall (CNGB accession - GWHBEBT00000000) and Chinese Dwarf (CNGB accession - GWHBEBU00000000) coconut genome assemblies (Fig. 6). This high-quality alignment, with no evident chromosomal rearrangements, underscores the robustness and quality of CGD genome assembly generated in the current study.

'Kalpa Genome Resource': modules and interface

Figure 7 gives an overview of the architecture of the Kalpa Genome Resource web server. The information in this customized database has been grouped into five basic functional web pages: 'Home,' 'Annotation,' 'Genome Browser,' 'Transcriptome,' and 'Download' (Fig. 8).

(i) The 'Home' page holds descriptions about the database, genome statistics, and news updates (Fig. 8A).
(ii) The 'Annotation' page has multiple search functions: Locus Search, KEGG Search, GO Slim Search, Text Search, and BLAST Search for Sequence homology search against genome and transcriptome sequences. BLAST Search is achieved by using the NCBI-BLAST API. The "Annotation" page further has subpages, namely "Simple Sequence Repeats" (SSR) and "Repeat Elements." As the page name suggests, these pages provide details of identified SSRs and various repeat elements in the KGB database (Fig. 8B and C).
(iii) The 'Genome Browser' page allows users to navigate through genomic sequences and their genomics elements interactively. The page also provides for the downloading of sequences for interested targeted regions or selected genomics elements. However, this feature is restricted to users and controlled by the admin. The architecture of this page also helps us integrate any genomic version in the future if the need arises. The 'Genome Browser' page integrates JBrowse for its functionality (Fig. 8D).

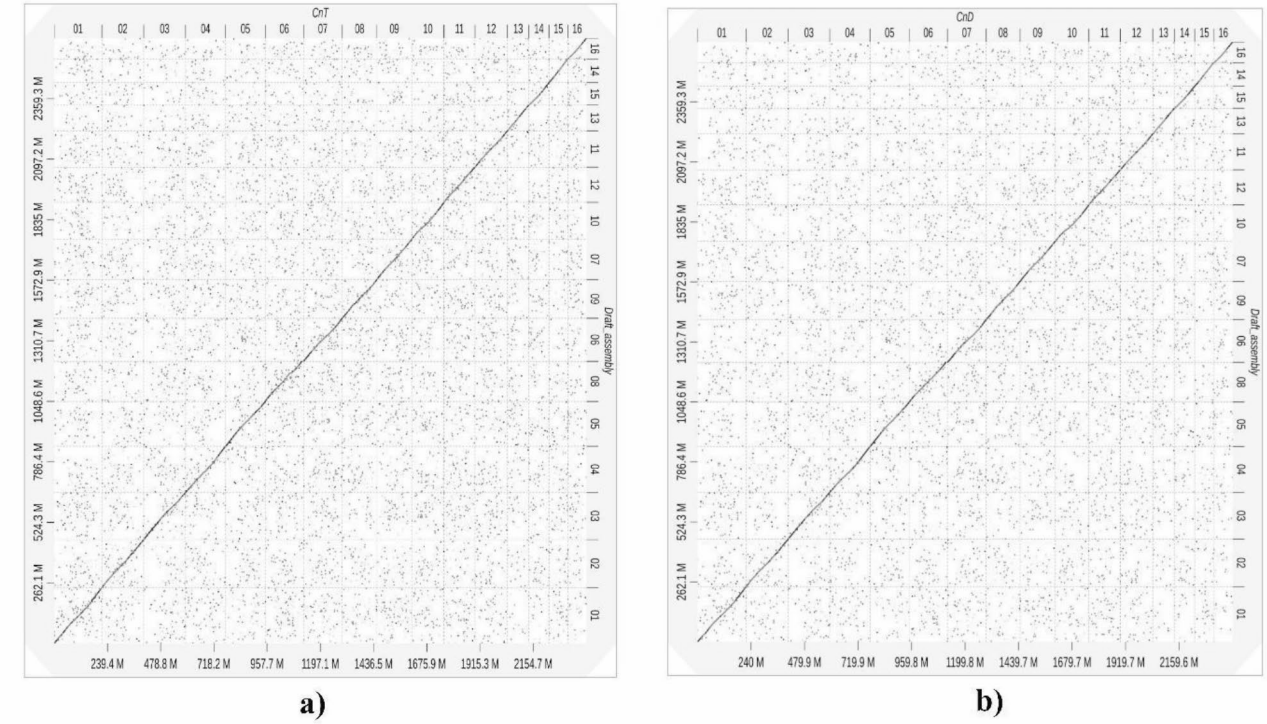


Fig. 6. Syntenic dot plot between Chowghat Green Dwarf (CGD) coconut cultivar and (**A**) Chinese Tall and (**B**) Chinese Dwarf coconut cultivars. The dot plot axis matrix is expressed in nucleotides, with the dot plot axes exhibiting a square relationship.

(iv) The 'Transcriptome' page contains an overview of the predicted transcripts from the genome and their comprehensive annotation. This database's architecture allows users to update multiple genome or transcriptome data and metadata as required without changing the existing database structure or functionality.

Search modules of the 'Kalpa Genome Resource' browser

This section describes some common queries that can be addressed using the 'Kalpa Genome Resource' browser. Detailed instructions for carrying out these and other queries can be found at Kalpa Genome Resource.

The simplest query search methods available are:

- **Locus search**: Find an object with a known KGR locus number.
- **GOSlim search**: Find objects associated with a known sequence database accession number.
- **KEGG search**: Find objects by KEGG Pathways Ids.
- **Text search**: Find objects that contain one or more keywords anywhere in their text.
- **BLAST search**: Sequence homology search against Genome & Transcriptome database.

Locus search

The Locus search function provides the option to search the KGR database with one locus ID at once. The search function returns the functional details of the locus searched. The details contain information like (i) Genome Version, (ii) Gene product Name, and (iii) Comprehensive Gene Ontology (GO) categorized as Biological Process, Molecular Function, & Cellular Component. It also contains information about associated KEGG pathways in which the gene predicted plays a role. Furthermore, the details of sequence homology are provided based on the identified protein ID from the NCBI RefSeq protein database.

GOSLIM search

The GOSLIM search function provides the option to search the KGR database with plant GOSLIM IDs and their annotation. These plant-specific GOSLIM IDs are associated with Locus ID during their sequence homology-based annotation process. The search page provides two options to users. They can search using the 'Locus ID'

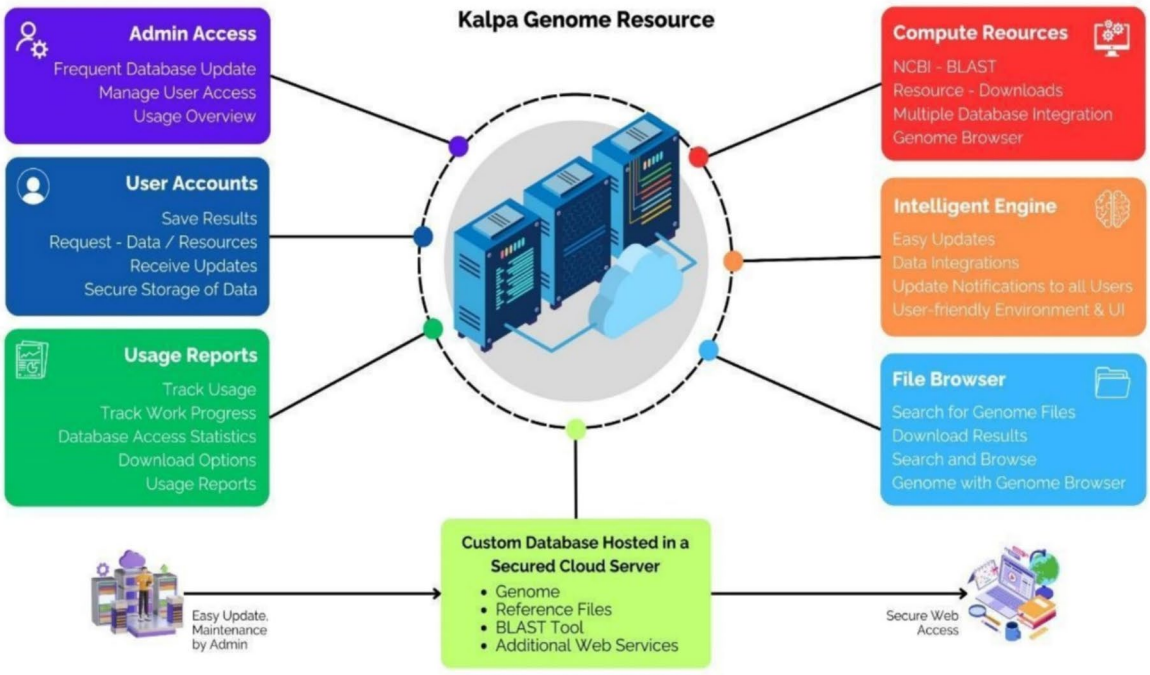


Fig. 7. Overview of the 'Kalpa Genome Resource' web server architecture.

or the 'GOSLIM ID' of their interest. The locus-based search function returns the list of GOSLIM IDs and their details associated with the Locus ID. However, the GOSLIM ID-based search returns the list of all the Locus IDs in the KGR database, which are reported to be associated with the GOSLIM ID of interest with their annotation. The Locus IDs are hyperlinked to their detailed functional page in both search results. The search result page also includes a sub-search box to filter the results.

KEGG search
The KEGG search function provides the option to search the KGB database with KEGG Pathway IDs. These specific KEGG pathway IDs were associated with Locus ID during their sequence homology-based annotation process. The search page provides options to users to navigate through the KGB database to identify all the Locus IDs, which were annotated with the specific KEGG pathway ID of their interest. The search function returns the list of Locus IDs, their KEGG Gene Ortholog ID, and all the associated KEGG Pathways with respective Locus IDs. The search result page also hyperlinks to the publicly available KEGG Gene Orthology page for further detailed information. The search result page also includes a sub-search box to filter the results.

TEXT search
Text Search is the most powerful feature of the KGB Database Annotation search. It allows users to search or query the database with any keyword of interest. For example, if a user wants to identify all the Locus / Gene IDs annotated as "Receptor" proteins, they must type in the keyword in the text search box and initiate the search function. The search result page lists all the Locus IDs annotated with the word "Receptor". The result page also provides a hyperlink to all the Locus IDs and their detailed functional annotation page so users can get more information about the specific Locus ID. The result page further provides a sub-search box to filter the results with another text word to narrow down the results.

BLAST search
The BLAST search page contains options for entering a query sequence (genome or transcriptome sequence) by typing or uploading a FASTA file. The user can also select different BLAST parameters and algorithms. To achieve this, we have implemented the available NCBI-BLAST API. However, downloading database sequences that give hits with the query sequence is restricted, while viewing sequence alignment in a graphical format is free (Fig. 8E).

Querying by region of interest

A region of interest can be specified using a pair of flanking markers: genes, genomic coordinates, transcripts, amplifiers, or any other mapped object. Given a region of interest, the comprehensive map is searched to find all loci within it. These loci can be displayed in a table or graphically as slices through the comprehensive map or as

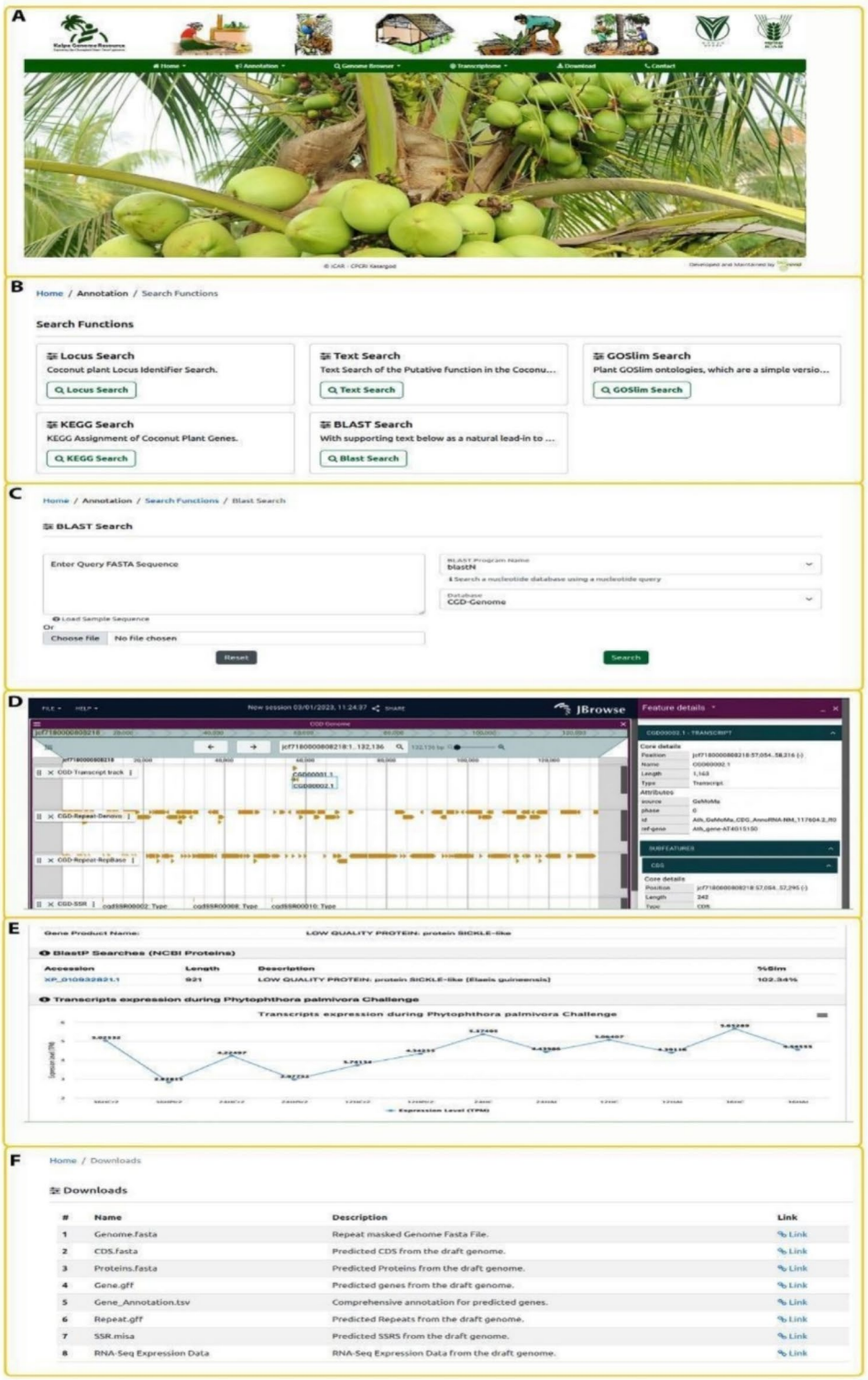


Fig. 8. Different pages of the 'Kalpa Genome Resource' web server. (**A**) Home page; (**B**) Different search functions; (**C**) BLAST search option; (**D**) JBrowse gene visualization; (**E**) BLASTp search results; and (**F**) Download options.

slices through a chosen set of primary maps. The comprehensive map slice shows all loci in the region, including genes, repeats, SSRs, etc.

Transcriptome datasets

The KGR database also stores and displays functional genomics and transcriptomic datasets, e.g., Gene Ontologies (GO), pathways, expression data, etc. This makes it easier to understand biological processes under normal or treated conditions. Researchers trying to identify similarities and differences between molecular conditions can upload tissue expression data. Expression profiles of the transcriptome datasets can also be visualized in graphical format and retrieved through the KGR database.

Retrieving a graphical view of the locus position

The results of queries for genes, primers, ESTs, etc., can be displayed on the KGR comprehensive map. Users can retrieve various default track information and customized track information like SNP, InDels, SSRs, repeats, and transcripts from the core database. If the results are spread across several chromosomes, multiple chromosomes will be displayed. Further, double-clicking on any of these genes reveals detailed information for the selected gene.

Jbrowse module of KGR

JBrowse is a popular interactive browser for visualizing genome data, encompassing genome sequences, gene structures, protein-coding gene annotations, single nucleotide polymorphisms (SNPs), site information, and expression profiles. This study incorporated the JBrowse2 tool into the KGR, importing the CGD genome sequence generated in the current study, RNA-Seq, and annotation data (in 'general feature format' files) into JBrowse (Fig. 8A).

Downloads

The portal provides a download option for accessing the CGD genome and transcriptome data in various file formats. All the generated files and images can be extracted from the 'Home page' by clicking the link using the download button (Fig. 8F).

In summary, we have assembled a high-quality genome of the *C. nucifera* CGD cultivar by incorporating ONT long-read sequencing, highly accurate short-read sequencing, and Hi-C technologies. The genome assembly generated and its accompanying resources are a significant enhancement for the palm genomics community, in general, and the coconut genomics community, in particular. They will provide valuable tools to aid in coconut breeding and deepen our understanding of the biology and evolution of this important tropical palm. As new and diverse omics data from the multi-omics platform (viz., genomics, transcriptomics, proteomics, metabolomics) become available in coconut, e.g., from re-sequencing, pan-genomes, and epigenome experiments, the 'Kalpa Genome Resource' will continually evolve. Future development will prioritize updating, integrating, and centralizing information and incorporating advanced functionalities to facilitate easier and more comprehensive access to this genomic resource. As more omics data for coconuts become available, the Kalpa Genome Resource will support the coconut research community and researchers and breeders of other palms.

Data availability

The raw DNA sequencing data from the *Cocos nucifera* Chowghat Green Dwarf cultivar project have been deposited at NCBI under the BioProject accession number PRJNA413280 and BioSample accession numbers SAMN42863343 (ONT sequencing), SAMN42863344 (Illumina sequencing), and SAMN42863345 (Hi-C). The genome assembly of the same has been deposited and can be accessed with accession number PDMH00000000 on NCBI. All the supplementary data files generated in this study are available in the Figshare Repository (https://figshare.com/s/464eb2f8212cd70f1ab5).The 'Kalpa Genome Resource' web server is freely accessible at http://210.89.54.198:3000/.

Received: 21 August 2024; Accepted: 12 November 2024
Published online: 20 November 2024

References

1. Arunachalam, V. & Rajesh, M. K. Coconut genetic diversity, conservation and utilization. In *Biodiversity and Conservation of Woody Plants* (eds. Ahuja, M. R. & Jain, S. M.) 3–36 (Springer, 2017).
2. Perera, L., Baudouin, L. & Mackay, I. SSR markers indicate a common origin of self-pollinating dwarf coconut in South-East Asia under domestication. *Sci. Hortic.* **211**, 255–262 (2016).
3. International Coconut Community (ICC). (2023). https://statistics.coconutcommunity.org/.
4. Jayasekhar, S. & Chandran, K. P. World economic importance. In *The Coconut Genome* 1–12 (Springer, 2021).
5. Yang, Y. et al. Coconut genome assembly enables evolutionary analysis of palms and highlights signalling pathways involved in salt tolerance. *Commun. Biol.* **4**, 105 (2021).
6. Wang, S. et al. High-quality reference genome sequences of two coconut cultivars provide insights into evolution of monocot chromosomes and differentiation of fiber content and plant height. *Genome Biol.* **22**, 1–25 (2021).
7. Shapiro, J. A. Engines of innovation: Biological origins of genome evolution. *Biol. J. Linn. Soc.* **139**, 441–456 (2023).
8. Gunn, B. F., Baudouin, L. & Olsen, K. M. Independent origins of cultivated coconut (*Cocos nucifera* L.) in the old world tropics. *PLOS ONE*. **6**, e21143 (2011).
9. Rajesh, M. 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).
10. Rajesh, M. K., Ramesh, S. V., Karun, A. & &Chowdappa, P. Genome sequencing, transcriptomics, proteomics and metabolomics. In *The Coconut Genome* (eds. Rajesh, M. K., Ramesh, S. V., Perera, L. & Kole, C.) 119–132 (Springer, 2021).

11. Rajesh, M. K. et al. Identification of expressed resistance gene analog sequences in coconut leaf transcriptome and their evolutionary analysis. *Turk. J. Agric. For.* **39**, 489–502 (2015).
12. Xiao, Y. et al. The genome draft of coconut (*Cocos nucifera*). *Gigascience* **6**, gix095 (2017).
13. Santhi, C. K. et al. Genome-wide exploration of auxin response factors (ARFs) and their expression dynamics in response to abiotic stresses and growth regulators in coconut (*Cocos nucifera* L). *Plant. Gene*. **28**, 100344 (2021).
14. Saensuk, C. et al. De novo transcriptome assembly and identification of the gene conferring a pandan-like aroma in coconut (*Cocos nucifera* L). *Plant. Sci.* **252**, 324–334 (2016).
15. Manohar, A. N. et al. Genome-guided molecular characterization of oil genes in coconut (*Cocos nucifera* L). *Philipp J. Sci.* **148** (S1), 153–164 (2019).
16. Xia, W. et al. Alternative splicing of flowering time gene FT is associated with halving of time to flowering in coconut. *Sci. Rep.* **10**, 11640 (2020).
17. Caro, R. E. et al. Mining and validation of novel simple sequence repeat (SSR) markers derived from coconut (*Cocos nucifera* L.) genome assembly. *J. Genet. Eng. Biotechnol.* **20**, 71 (2022).
18. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).
19. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics*. **29**, 2669–2677 (2013).
20. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **19**, 1 (2018).
21. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
22. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: Yet another Hi-C scaffolding tool. *Bioinformatics*. **39**, btac808 (2023).
23. Wingett, S. et al. HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Res*. **4**, 1310 (2015).
24. 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).
25. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).
26. Rhie, A. et al. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
27. 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).
28. Flynn, J. et al. (ed, M.) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117** 9451–9457 (2020).
29. Bao, W. & Kojima, K. K. &Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*. **6**, 1–6 (2015).
30. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* **5**, 4–10 (2004).
31. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18 (2019).
32. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 491 (2011).
33. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinf.* **2**, lqaa026 (2020).
34. Stanke, M. et al. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
35. Rajesh, M. K. et al. De novo assembly and characterization of global transcriptome of coconut palm (*Cocos nucifera* L.) embryogenic calli using Illumina paired-end sequencing. *Protoplasma*. **253**, 913–928 (2016).
36. Venugopal, G. et al. Transcriptome assembly of coconut endosperm callus (Laccadive Micro Tall Cv.) And its functional annotation. *J. Plantn Crops*. **49**, 225–230 (2021).
37. Rajesh, M. K. et al. Comparative transcriptome profiling of healthy and diseased Chowghat Green Dwarf coconut palms from root (wilt) disease hot spots. *Eur. J. Plant. Pathol.* **151**, 173–193 (2018).
38. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
39. Jones, P. et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*. **30**, 1236–1240 (2014).
40. Törönen, P., Medlar, A. & Holm, L. PANNZER2: A rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).
41. Kanehisa, M. KEGG GLYCAN. *A Practical Guide to Using Glycomics Databases* **6**, 177–193 (2016).
42. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
43. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013).
44. Griffiths-Jones, S. Annotating non-coding RNAs with Rfam. *Curr. Protoc. Bioinf.* **9**, 12–15 (2005).
45. Calle Garcia, J. et al. PRGdb 4.0: An updated database dedicated to genes involved in plant disease resistance process. *Nucleic Acids Res.* **50**, D1483–D1490 (2022).
46. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20** (4), 1160–1166 (2019).
47. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37.5**, 1530–1534 (2020).
48. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, gkae268 (2024).
49. Chen, H., Zwaenepoel, A. & Van de Peer, Y. Wgd v2: A suite of tools to uncover and date ancient polyploidy and whole-genome duplication. *Bioinformatics*. **40**, btae272 (2024).
50. Wang, Y. et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
51. Cabanettes, F. & Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. **6**, e4958 (2018).
52. Gangaraj, K. P. & Rajesh, M. K. Dataset of dual RNA-sequencing of *Phytophthora palmivora* infecting coconut (*Cocos nucifera* L). *Data Brief.* **30**, 105455 (2020).
53. Kim, D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
54. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
55. Gertz, E. M. et al. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 1–4 (2006).
56. Al-Mssallem, I. S. et al. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun.* **4**, 2274 (2013).
57. Singh, R. et al. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature*. **500**, 335–339 (2013).
58. Schley, R. J. et al. The ecology of palm genomes: Repeat-associated genome size expansion is constrained by aridity. *New. Phytol.* **236**, 433–446 (2022).

59. Lantican, D. V. et al. De novo genome sequence assembly of dwarf coconut (*Cocos nucifera* L.'Catigan Green Dwarf') provides insights into genomic variation between coconut types and related palm species. *G3: Genes, Genomes, Genetics* **9**(8), 2377–2393 (2019).
60. Shah, S. H., Islam, S., Mohammad, F. & Siddiqui, M. H. Gibberellic Acid: A versatile regulator of plant growth, development and stress responses. *J. Plant Growth Regul.* **42**, 7352–7373 (2023).
61. Rahman, S., Gangaraj, K. P., Muralikrishna, K. S. & Rajesh, M. K. Isolation and characterisation of full-length coding sequences of gibberellic acid biosynthetic genes in coconut (*Cocos nucifera* L.) and their expression analysis. *South. Afr. J. Bot.* **153**, 297–307 (2023).
62. Zhou, L. et al. Regulation of oil biosynthesis and genetic improvement in plants: Advances and prospects. *Genes*. **15**, 1125 (2024).
63. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant. Cell.* **26**, 2792–2802 (2014).

Author contributions
R.M.K. conceived, designed, and managed the project; R.M.K., R.B., R.S., P.K.O., S.G., B.P., R.J.T., Ak.D., J.S., C.K.P., M.K.S., N.B.J. and Al.D. performed the experiments, and collected and analyzed the data; R.B., R.S., B.P., J.S. and C.K.P. constructed the database; R.M.K., R.B. and R.S. designed the layout of the web pages; M.K.R., R.B., R.S., P.K.O., S.G., B.P. drafted the manuscript. R.J.T., Ak.D., J.S., C.K.P., M.K.S., N.B.J. and Al.D. revised the manuscript. All authors read, commented on, and approved the manuscript.

Funding
The authors acknowledge the funding for this work from the Indian Council of Agricultural Research (ICAR-CPCRI Project No. 1000761030) and the ICAR-NEH fund of ICAR-CPCRI, Kahikuchi, India.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-79768-3.

Correspondence and requests for materials should be addressed to M.K.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.