

Chromosome-level genomes of seeded and seedless date plum based on third-generation DNA sequencing and Hi-C analysis

Weitao Mao^{1,2,3#}, Guoxin Yao^{2#}, Shangde Wang¹, Lei Zhou⁴, Guosong Chen⁵, Ningguang Dong^{1*} and Guanglong Hu^{1*}

¹ Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China), Ministry of Agriculture, Beijing Engineering Research Center for Deciduous Fruit Trees, Beijing Academy of Forestry and Pomology Sciences, Beijing 100093, China

² Hubei Key Laboratory of Quality Control of Characteristic Fruits and Vegetables, College of Life Science and Technology, Hubei Engineering University, Xiaogan 432000, China

³ School of Life Science, Hubei University, Wuhan 430062, China

⁴ Hubei Key Laboratory of Food Crop Germplasm and Genetic Improvement, Food Crops Institute, Hubei Academy of Agricultural Sciences, Wuhan 430072, China

⁵ Beijing XinTaoYuan Commerce & Trading Co., Ltd., Beijing 101215, China

These authors contributed equally: Weitao Mao, Guoxin Yao

* Corresponding authors, E-mail: dongng@sina.com; hglcau@gmail.com

Abstract

Diospyros lotus L. (Date plum) is an important tree species that produces fruit with a high nutritional value. An accurate chromosomal assembly of a species facilitates research on chromosomal evolution and functional gene mapping. In this study, we assembled the first chromosome-level genomes of seeded and seedless *D. lotus* using Illumina short reads, PacBio long reads, and Hi-C technology. The assembled genomes comprising 15 chromosomes were 617.66 and 647.31 Mb in size, with a scaffold N50 of 40.72 and 42.67 Mb for the seedless and seeded *D. lotus*, respectively. A BUSCO analysis revealed that the seedless and seeded *D. lotus* genomes were 91.53% and 91.60% complete, respectively. Additionally, 20,689 (95.4%) and 22,844 (98.5%) protein-coding genes in the seedless and seeded *D. lotus* genomes were annotated, respectively. Comparisons of the chromosomes between genomes revealed inversions and translocations on chromosome 8 and inversions on chromosome 11. We identified 490 and 424 gene families that expanded in the seedless and seeded *D. lotus*, respectively. The enriched pathways among these gene families included the estrogen signaling pathway, the MAPK signaling pathway, and biosynthetic pathways for flavonoids, monoterpenoids, and glucosinolates. Moreover, we constructed the first *Diospyros* genome database (<http://www.persimmongenome.cn>). On the basis of our data, we developed the first high-quality annotated *D. lotus* reference genomes, which will be useful for genomic studies on persimmon and for clarifying the molecular mechanisms underlying important traits. Comparisons between the seeded and seedless *D. lotus* genomes may also elucidate the molecular basis of seedlessness.

Citation: Mao W, Yao G, Wang S, Zhou L, Chen G, et al. 2021. Chromosome-level genomes of seeded and seedless date plum based on third-generation DNA sequencing and Hi-C analysis. *Forestry Research* 1:9 <https://doi.org/10.48130/FR-2021-0009>

INTRODUCTION

Date plum (*Diospyros lotus* L.), which belongs to the genus *Diospyros* in the family Ebenaceae, is an important deciduous fruit tree species that grows in Asia, where it is cultivated for its edible fruit. The *Diospyros* genus, within the Ebenaceae (Ericales), contains more than 700 species, including the economically important persimmons (*D. kaki*, *D. virginiana*, and *D. lotus*) and ebony (*D. ebenum*)^[1–3]. The fruit of *D. lotus* is globe shaped and yellow or bluish-black when mature^[4,5]. Able to be grown at 2,200 m above sea level, *D. lotus* is the most cold-tolerant *Diospyros* species in China. It is used as a rootstock because of its high grafting capability and for developing new varieties because of its strong tolerance to drought and cold^[6]. Additionally, *D. lotus* is used in drug research. The *D. lotus* fruit is used as a sedative, astringent, food and laxative, and has antiseptic, antidiabetic, antitumor, and antipyretic properties. It is also useful for treating constipation and diarrhea, dry coughs and hypertension^[7].

There are seeded and seedless *D. lotus* varieties, and the seedless type has a high nutritional value. The edible parts are ideal raw materials for research and the development of foods, drinks and health-care products^[8]. Many high-quality fruits are unacceptable for consumers because they have too many seeds or their seeds are too large^[9]. The production of seedless fruit is also attractive because it avoids the possibility of any undesirable pollination. Seedless fruit is an important and peculiar horticultural trait that has been selected for and retained during long-term cultivation. The seedless trait of fruit is very complex, and is not only affected by internal genetic factors in certain tree species and varieties, but also by external factors. Seedless fruit can be obtained using specific treatments. The addition of a certain CuSO₄·5H₂O concentration while cross-pollinating during the citrus flowering period significantly reduces the numbers of seeds in the fruit without affecting yield^[10]. Similarly, watermelon fruit that results from pollination with pollen irradiated with soft-X-ray contains only empty seed, although the fruit

develops to a normal size^[11]. Additionally, gibberellic acid treatments induce parthenocarpy in Algerie loquat^[12].

To date, most of the research on the mechanisms underlying seedlessness has been at the cellular level, and the application of gene sequencing technology has revealed that the expression of certain genes causes fruit to be seedless. To the best of our knowledge, there are no published reports on the seedless trait of *D. lotus* fruit. Therefore, the aim of this study was to provide new insights into the production of seedless *D. lotus* fruit through genome sequencing and a comparative analysis of seeded and seedless *D. lotus* varieties. The results will also be useful for breeding *D. lotus* varieties with desirable characteristics.

MATERIALS AND METHODS

Plant material and DNA sequencing

Two *D. lotus* varieties (Fig. 1), seedless *D. lotus* (W01) and seeded *D. lotus* (Yz01), were grown in Taoyuan Village, Zhenluoying Town, Pinggu District, Beijing, China. Fresh, healthy leaves were collected and immediately frozen in liquid nitrogen. Genomic DNA extracted from the samples using the cetyltrimethylammonium bromide method^[13] was used for sequencing. To obtain sufficient high-quality DNA for the PacBio Sequel II platform (Pacific Biosciences of California Inc., Menlo Park, CA, USA), the concentration and purity of the extracted DNA were determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit fluorometer (Thermo Fisher Scientific). Moreover, the integrity of the DNA was checked by 1% agarose gel electrophoresis. The extracted DNA was sequenced on the Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA) and PacBio Sequel II platforms (Pacific Biosciences of California Inc., Menlo Park, CA, USA). The short reads generated from the Illumina platform were used to estimate the genome size, heterozygosity, and repeat content, whereas the long reads from the PacBio platform were used for assembling genomes. Briefly, qualified DNA samples were randomly fragmented into 350 base pair (bp) segments using ultrasonic crushing apparatus, after which they were used for the end repair, poly (A) addition, barcode indexing, purification, and PCR amplification steps. Regarding the Illumina NovaSeq sequencing analysis, we constructed a paired-end library with 150 bp sequences using the manufacturer-recommended method. After filtering, 80.99 Gb (119.53-fold genome sequence coverage) and 79.21 Gb (114.98-fold genome sequence coverage) of clean data were generated for the seedless and seeded *D. lotus*, respectively. For the PacBio sequencing, SMRTbell libraries (approximately

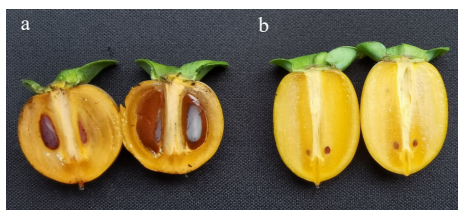


Fig. 1 Fruit shapes and seed sizes of the seeded and seedless *Diospyros lotus*. (a) Seeded (Yz01); (b) Seedless (W01).

20 kb) were obtained according to the PacBio protocol. After removing adapters and correcting and trimming the data, 92.08 Gb (103.29-fold genome sequence coverage) and 133.51 Gb (166.98-fold genome sequence coverage) of sequence data were generated for the seedless and seeded *D. lotus*, respectively.

RNA extraction and sequencing

Total RNA was prepared from the two *D. lotus* varieties stems, fruits and leaves using TRIzol reagent (Invitrogen, California, USA). A NanoDrop 2000 spectrophotometer (Waltham, MA, USA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) were applied to check RNA quality. Equal amounts of RNA from each tissue were used for cDNA library construction. Approximately 4.17 Gb and 4.15 Gb of transcript data were produced for seedless and seeded *D. lotus* from the Illumina HiSeq X Ten sequencing platform and processed using Trimmomatic (version 0.36) with the default parameters.

Genome *de novo* assembly

Genomes were assembled using Canu (version 1.5)^[14], with the following parameters: maxThreads = 200, minReadLength = 1,000, corOut Coverage = 40, correctedErrorRate = 0.045, minOverlapLength = 500, rawErrorRate = 0.3, and corMin Coverage = 4. To increase the accuracy of the sequencing data, the genomes were assembled using the error correction, trimming, and assembly steps of Canu. The sequences were polished in two rounds. Specifically, the PacBio long-read sequence data were polished with Arrow^[15], after which the Illumina sequence data were polished using Pilon (version 1.22)^[16]. Purge Haplotigs^[17] was used to remove genomic redundancies after the initial assembly and correction.

K-mer analysis of the *D. lotus* genomes

We used Illumina short reads and a k-mer-based method^[18] to estimate the size, heterozygosity, and repeat content of the seedless and seeded *D. lotus* genomes, using a software package (GCE-1.0.2, <https://github.com/fanagislab/GCE>). The k-mer frequency (k = 17) was determined using Jellyfish software, and the frequency distribution derived from the sequencing reads was plotted.

High-quality assembly using Hi-C technology

Fresh young *D. lotus* leaves were treated with paraformaldehyde. Chromatin was digested with the restriction enzyme *Mbo*I and ligated *in situ* after a biotinylation step. The 5' overhangs were labeled with a biotinylated tag and repaired. Following the ligation, the DNA was extracted and sheared, after which fragments between 300 and 500 bp long were selected. The biotin-containing fragments were captured to construct a library, which was then sequenced with the Illumina system. The Hi-C library sequencing for the seedless and seeded *D. lotus* resulted in 86.96 Gb and 107.87 Gb data, respectively (Table 1). The two groups of sequencing reads were aligned to the previously assembled genomes using Bowtie2^[19]. The Hi-C data were identified and aligned, and the repeated reads were removed. The data were filtered and evaluated in tandem using HiCUP^[20]. On the basis of cis interactions, rather than trans interactions, contigs or scaffolds were divided, anchored, sequenced, directed and

Table 1. Summary of the sequencing data used for assembling the *Diospyros lotus* genomes.

Library type	Seedless <i>Diospyros lotus</i> (W01)			Seeded <i>Diospyros lotus</i> (Yz01)		
	Library size (bp)	Clean data (Gb)	Coverage (x)	Library size (bp)	Clean data (Gb)	Coverage (x)
Illumina	350	80.99	119.53	350	79.21	114.98
Pacbio	20,000	92.1	103.29	20,000	133.51	166.98
Hi-C	350	86.96	–	350	107.87	–

incorporated to obtain chromosome-level genomes using 3D-DNA^[21].

Genome assembly quality evaluation

To evaluate the quality of the assembled genomes, the Illumina short reads were mapped to the genomes using the BWA software^[22] and the PacBio long reads were mapped using BLASR^[23]. The completeness of the assembled genomes was determined by BUSCO analyses^[24] using the actinopterygii_odb9 dataset. Long terminal repeat (LTR) sequences were used to evaluate genomic integrity, which was expressed as the LTR assembly index (LAI), using the LTR_finder and LTR_retriever programs^[25]. Illumina short reads were aligned to the genome using SAMtools^[26], whereas Picard tools^[27] were used to detect mutations and GATK^[28] was used to count the homozygous and heterozygous SNPs and InDels. The results are herein presented as circular genomic maps.

Genome annotation

Repetitive sequences, including transposable elements (TEs) and tandem repeats, were analyzed. More specifically, the repeated sequences in the *D. lotus* genomes were annotated using homology-based and *ab initio* prediction methods. RepeatMasker and Repeat Protein Mask (version 4.0.5)^[29] were used to retrieve data from the RepBase database (<http://www.girinst.org/repbase>). Tandem Repeats Finder^[30] and LTR_finder were used to make *ab initio* predictions.

The protein-coding genes were annotated using a combination of homology-based, *ab initio*, and transcriptome-based predictions. Augustus (version 3.0.2)^[31] was used to predict *ab initio* coding genes. For the homology-based method, protein sequences from related plants, including *Olea europaea*, *Capsicum annuum*, *Daucus carota*, *Solanum pennellii*, *Arabidopsis thaliana*, *Lactuca sativa*, and *Solanum tuberosum*, were downloaded from public databases and aligned against the *D. lotus* genomes using TBLASTN (E-value < 1e-5)^[32]. The sequences derived from RNA-seq data were compared with the assembled *D. lotus* genomes to identify potential exon regions using TopHat (version 2.0.8)^[33] and Cufflinks (version 2.1.1)^[34]. We integrated all predicted genes using MAKER software^[35]. The following databases were screened for homologous sequences: NCBI non-redundant protein (NR), Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Eukaryotic Orthologous Groups (KOG), SwissProt, TrEMBL, InterProScan, and Pfam.

The default parameters of tRNAscan-SE^[36] were used to predict transfer RNA (tRNA) genes. Because ribosomal RNAs (rRNAs) are highly conserved, the rRNA sequences of related species were selected as reference sequences and used to search for rRNA sequences in the genomes via a BLASTN alignment (E-value < 1e-5). The microRNA (miRNA) and small nuclear RNA (snRNA) fragments were identified by searching

the Rfam database (version 11.0)^[37] using INFERNAL (version 1.1)^[38].

Synteny analysis of the seedless and seeded *D. lotus* genomes

The evolution of the seedless and seeded *D. lotus* chromosomes as well as gene synteny were investigated using MCScan^[39]. A total of 17,162 gene pairs were detected in the comparisons of the seedless and seeded *D. lotus* genomes. The aligned syntenic chromosomes were visualized.

Analysis of genome evolution

To more thoroughly examine the phylogenetic relationships of *D. lotus* and the evolution of its gene families, we clustered gene sequences from 17 related plant species and performed a phylogenetic analysis based on the protein-coding genes from the seedless and seeded *D. lotus* and the 17 other species. We extracted and downloaded the protein sequences encoded by single-copy genes from the NCBI database for the following 17 species: *Malus domestica*, *Citrus reticulata*, *Juglans regia*, *Solanum lycopersicum*, *Diospyros oleifera* Cheng, *Rhododendron delavayi*, *Camellia sinensis*, *Coffea canephora*, *Daucus carota*, *Coriandrum sativum*, *Cucurbita pepo*, *Vitis vinifera*, *Eriobotrya japonica*, *Sorghum bicolor*, *Arabidopsis thaliana*, *Oryza sativa* subsp. *japonica*, and *Beta vulgaris*. Analyses were conducted using BLASTP (E-value < 1e-5)^[40] and OrthoFinder (version 2.27)^[41], with an inflation parameter of 1.5. To reveal the phylogenetic relationships among *D. lotus* and the other species, the protein sequences encoded by single-copy orthologous genes were aligned using MUSCLE (version 3.8.31)^[42]. These phylogenetic analyses were performed according to the maximum-likelihood method of PhyML (version 3.0)^[43]. Using the molecular clock data from the TimeTree database, the divergence times were determined with the approximate likelihood calculation method of PAML (version 4.8)^[44]. We compared the cluster size differences between the ancestors and each species and analyzed the expansion and contraction of the gene families using CAFE (version 2.1)^[45].

Construction of the *Diospyros* genome database

The *Diospyros* Genome Database was set up using Tomcat and MySQL. The backend was designed and implemented using the SpringBoot + MyBatis framework, with CentOS as the server. Data were visualized using an open source ECharts package. We collected genomic data for *Diospyros oleifera* Cheng, *Diospyros lotus*_Kunsenshi-Male, seedless *Diospyros lotus* and seeded *Diospyros lotus*.

Data availability

The sequencing datasets and genome assemblies have been deposited in public repositories. The Illumina genome sequencing data were deposited in the NCBI Sequence Read Archive under the accession numbers SRR12450967

(Seedless) and SRR12450964 (Seeded). The PacBio genome sequencing data were deposited in the NCBI Sequence Read Archive under the accession numbers SRR12450966 (Seedless) and SRR12450963 (Seeded). The Hi-C sequencing data were deposited in the NCBI Sequence Read Archive under the accession numbers SRR12450965 (Seedless) and SRR12450962 (Seeded). The final chromosome assemblies were deposited in the NCBI GenBank database under accession numbers JACNMG000000000 (Seedless) and JACRTX000000000 (Seeded). Raw sequencing data for RNA-Seq used for annotation have been deposited in the NCBI under the SRA accession number SRR14028490 (Seeded) and SRR14026913 (Seedless).

RESULTS AND DISCUSSION

Genome sequencing and assembly

On the basis of the k-mer analysis (k-mer = 17), the seedless *D. lotus* genome size was 682 Mb, with a heterozygosity of 1.0% and a repeat content of 57.15%, whereas the seeded *D. lotus* genome size was 616 Mb, with a heterozygosity of 1.26% and a repeat content of 54.92%. A total of 92.08 Gb (103.29-fold genome sequence coverage) and 133.51 Gb (166.98-fold genome sequence coverage) of PacBio long reads, as well as 80.99 Gb (119.53-fold genome sequence coverage) and 79.21

Gb (114.98-fold genome sequence coverage) of Illumina clean data, were generated for the seedless and seeded *D. lotus*, respectively (Table 1). The total length of the assembled reads for the seedless *D. lotus* genome was 617.66 Mb, which included 706 contigs. The contig N50 was 3.01 Mb and the longest contig was 16.26 Mb. The total length of the assembled reads for the seeded *D. lotus* genome was 647.31 Mb, which included 743 contigs. The contig N50 was 2.46 Mb and the longest contig was 14.82 Mb (Table 2). The size difference between the final genomes and the genome survey sequences may have been because of the heterozygosity and repetitive sequence of the *D. lotus* genomes. On the basis of the Hi-C assisted assembly, 142 contigs were successfully clustered into 15 chromosomes in the seedless *D. lotus* genome, and the scaffold N50 reached 40.72 Mb (Supplemental Table S1), whereas 41 contigs were successfully clustered into 15 chromosomes in the seeded *D. lotus* genome, and the scaffold N50 reached 42.67 Mb (Supplemental Table S2). To the best of our knowledge, this is the first report of chromosome-level *D. lotus* genomes (Fig. 2).

Genome assembly quality evaluation

The Illumina short-read mapping rates were 97.74% and 98.24% for the seedless and seeded *D. lotus* genomes, respectively (Supplemental Table S3). The PacBio long-read mapping rates were 90.90% and 94.89% for the seedless and

Table 2. Summary of the assembled seedless and seeded *Diospyros lotus* genomes.

Parameter	Seedless <i>Diospyros lotus</i> (W01)		Seeded <i>Diospyros lotus</i> (Yz01)	
	Contig length (bp)	Contig number	Contig length (bp)	Contig number
N90	561,232	228	537,928	279
N80	1,144,354	151	1,078,450	194
N70	1,625,012	106	1,450,541	143
N60	2,258,638	73	2,059,392	106
N50	3,006,748	49	2,463,960	77
Total length	617,662,490	–	647,313,630	–
Number (≥ 100 bp)	–	706	–	743
Number (≥ 2 kb)	–	691	–	734
Max length	16,262,241	–	14,842,567	–

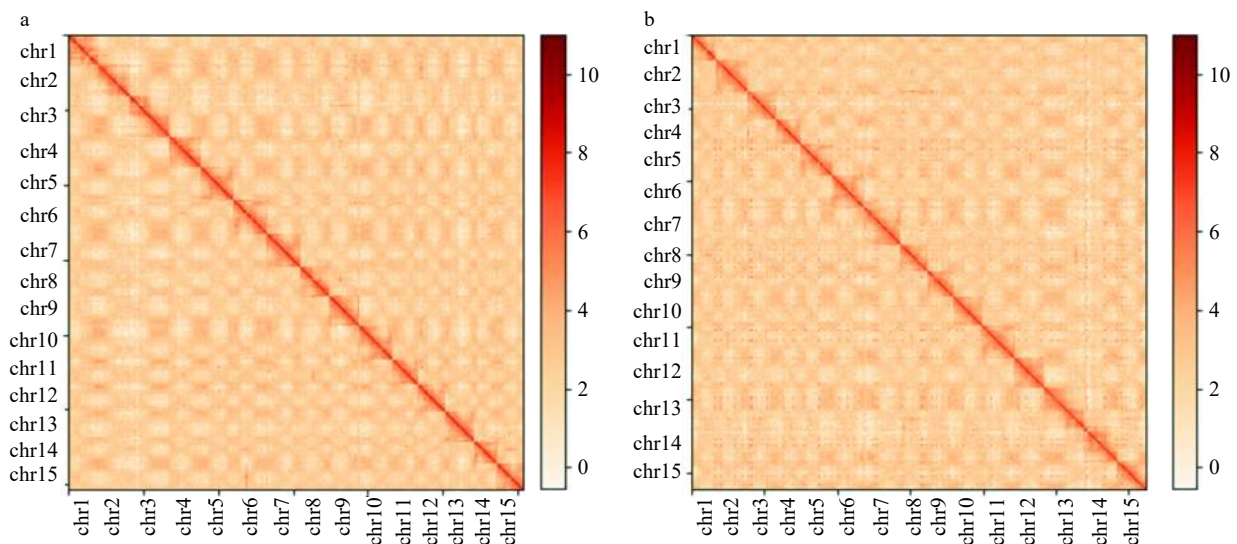


Fig. 2 Hi-C interaction heat maps for *Diospyros lotus* genomes presenting the interactions among 15 chromosomes. (a) Seeded (Yz01); (b) Seedless (W01).

seeded *D. lotus* genomes, respectively (Supplemental Table S4). The BUSCO analysis revealed that the seedless and seeded *D. lotus* genomes were 91.53% and 91.60% complete, respectively (Supplemental Table S5). The results of the analysis of the homozygosity and heterozygosity of the SNPs and InDels are presented in Supplemental Table S4. The LAI is a newly developed reference-free genome metric for evaluating genome assembly continuity using LTR retrotransposons. The LAI values for our assembled seedless *D. lotus* genome (LAI = 15.22) and seeded *D. lotus* genome (LAI = 11.98) were relatively high, exceeding the threshold for reference genome assemblies. Circular maps of the seedless and seeded *D. lotus* genomes are presented in Fig. 3.

Repetitive sequence annotation

A genomic analysis revealed that 69.56% of the seedless *D. lotus* genome consisted of repetitive sequences, of which TEs accounted for 68.76%. The most frequently detected TEs in the seedless *D. lotus* genome were LTR retrotransposons

(55.17%), followed by DNA TEs (10.20%) (Supplemental Table S7). In contrast, 73.81% of the seeded *D. lotus* genome consisted of repetitive sequences, of which TEs accounted for 72.87%. The most frequently detected TEs in the seeded *D. lotus* genome were LTR retrotransposons (59.37%), followed by DNA TEs (11.57%) (Supplemental Table S8).

Gene predictions and functional annotations

Homology-based, transcriptome-based, and *ab initio* gene predictions were used to generate gene models, which were combined. After eliminating sequence redundancies with MAKER, 21,684 and 23,193 protein-coding genes were identified in the seedless and seeded *D. lotus* genomes, respectively. The screening of the NR, GO, KEGG, KOG, SwissProt, TrEMBL, InterProScan, and Pfam databases for homologous sequences indicated that the seedless and seeded *D. lotus* genomes respectively contained 20,689 (95.41%) and 22,844 (98.50%) protein-coding genes listed in at least one public database (Table 3). The number of genes,

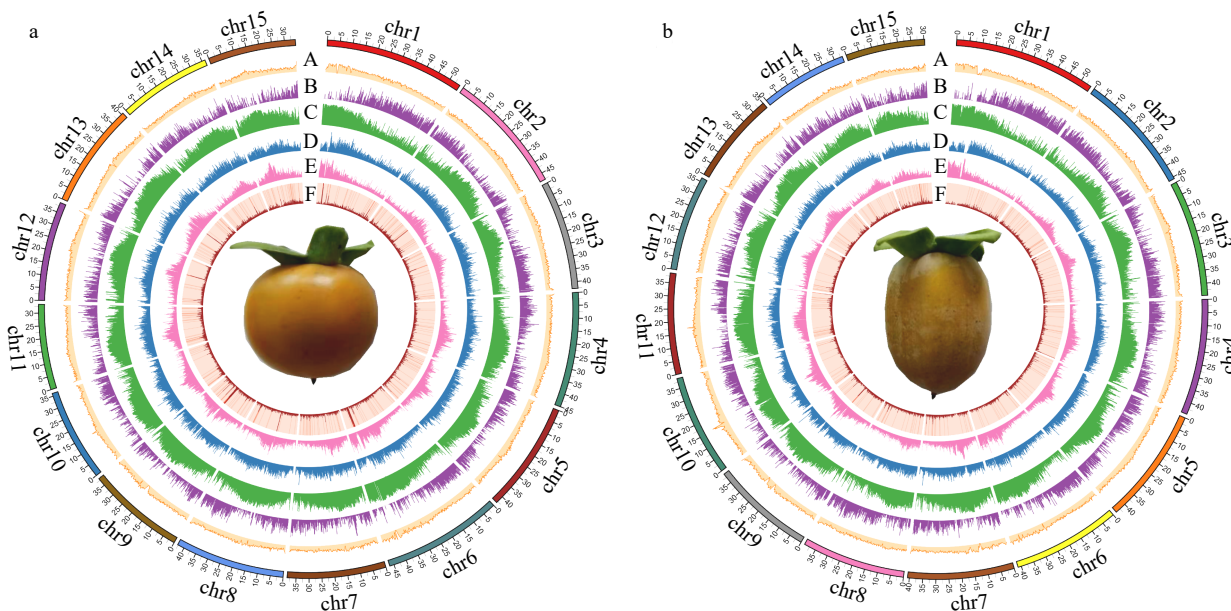


Fig. 3 Circular genomic maps for *Diospyros lotus*. (a) Seeded (Yz01); (b) Seedless (W01). A. GC content distribution; B. Gene density distribution; C. Repeats density distribution; D. LTR-Copia density distribution; E. LTR-Gypsy density distribution; F. DNA transposon density distribution.

Table 3. General statistics for the functional annotations of the genes in the seedless and seeded *Diospyros lotus* genomes.

Type	Seedless <i>Diospyros lotus</i> (W01)		Seeded <i>Diospyros lotus</i> (Yz01)		
	Number	Percent (%)	Number	Percent (%)	
Total	21,684	–	23,193	–	
Annotated	20,689	95.41	22,844	98.5	
	InterPro	17,473	80.58	20,037	86.39
	GO	12,161	56.08	14,066	60.65
	KEGG ALL	20,547	94.76	22,750	98.09
	KEGG KO	8,435	38.90	9,812	42.31
	Swissprot	15,057	69.44	16,896	72.85
	TrEMBL	20,587	94.94	22,790	98.26
	TF	1,560	7.19	1,572	6.78
	Pfam	17,064	78.69	19,709	84.98
	NR	20,607	95.03	22,794	98.28
KOG	17,990	82.96	20,208	87.13	
Unannotated	–	995	4.59	349	1.50

gene length distribution, coding sequence length distribution, exon length distribution, and intron length distribution for the *D. lotus* genomes were similar to the corresponding data for the other analyzed species (Supplemental Fig. S1, S2).

Noncoding RNA annotation

We identified snRNA, miRNA, and rRNA genes in the seedless and seeded *D. lotus* genomes based on a BLASTN search of the Rfam database (E-value < 1e-5), whereas we used tRNAscan-SE and RNAmmer to predict the tRNAs and rRNAs. Finally, 146 miRNAs, 496 tRNAs, 719 rRNAs, and 792 snRNAs were identified in seedless, with average lengths of 129, 75, 220, and 111 bp, respectively (Supplemental Table S9). Additionally, 219 miRNAs, 826 tRNAs, 2,386 rRNAs, and 1,371 snRNAs were identified in seeded, with average lengths of 127, 75, 376, and 110 bp, respectively (Supplemental Table S10).

Synteny analysis

We analyzed the synteny between the seedless and seeded *D. lotus* genomes, and the results are presented in Fig. 4b. Although the degree of synteny between the two genomes was relatively high, chromosome 8 included inversions and translocations and chromosome 11 contained inversions (Fig. 4a). These chromosomal variations may be related to differences in seedless traits, as has been reported for banana and citrus species^[9,46].

Evolution of the seedless and seeded *D. lotus* genomes

We selected genome sequences of representative plant species for a comparative genomic analysis of seedless and seeded *D. lotus* to reveal the genome evolution and divergence time of *D. lotus*. Seedless and seeded *D. lotus* and other 17 species were analysed together (Supplemental Table S11). A total of 8,998 gene families were shared by these five species, whereas 608 and 502 gene families were unique to seedless and seeded *D. lotus*, respectively (Fig. 5). There were significantly more unique gene families in seedless *D. lotus* than in seeded *D. lotus*. The phylogenetic analysis indicated that *D. lotus* is most closely related to *D. oleifera*, with an estimated divergence time of 23.5 million years. Seedless and seeded *D. lotus* were estimated to have diverged 5.9 million years ago (Fig. 6).

Expansion and contraction of gene families

The expansion and contraction of gene families are important processes during evolution^[47]. Our analysis indicated that 490 gene families expanded in seedless (Supplemental Fig. S3). The enriched KEGG pathways among these families included estrogen signaling pathway, MAPK signaling pathway, antigen processing and presentation, longevity regulating pathway—multiple species, flavonoid biosynthesis, and metabolism of xenobiotics by cytochrome P450 (Supplemental Table S12). The enriched GO terms included cargo receptor activity; electron transporter, transferring electrons within the cyclic electron transport pathway of photosynthesis activity; photosynthetic electron transport in photosystem II; and scavenger receptor activity (Supplemental Table S13). Our results indicated that 1,497 gene families contracted in seedless. Fatty acid elongation and amino sugar and nucleotide sugar metabolism were two

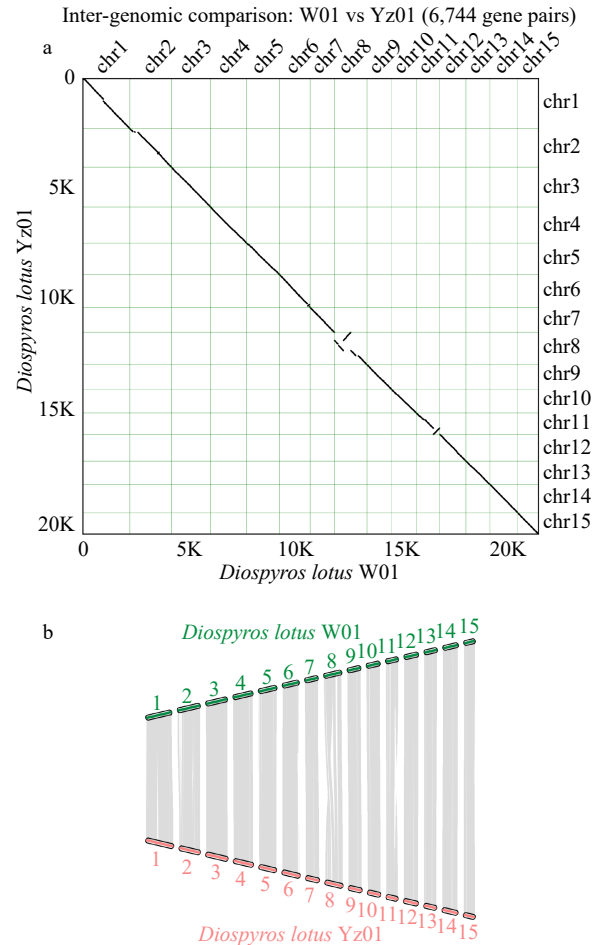


Fig. 4 Chromosomal synteny between seedless and seeded *Diospyros lotus*. (a) Inter-genomic comparison. (b) Chromosomal maps of seedless and seeded *Diospyros lotus*.

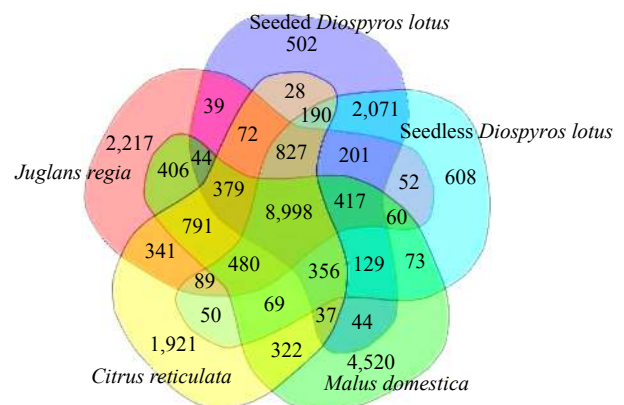


Fig. 5 Venn diagram presenting the number of shared and unique protein-coding genes among seedless and seeded *Diospyros lotus*, *Malus domestica*, *Citrus reticulata*, and *Juglans regia* revealed by an orthology analysis.

of the enriched KEGG pathways among these gene families (Supplemental Table S14). The most enriched GO term was catalytic activity (Supplemental Table S15). In contrast, our analyses indicated that 424 gene families expanded in seeded. The functional annotation of these genes revealed

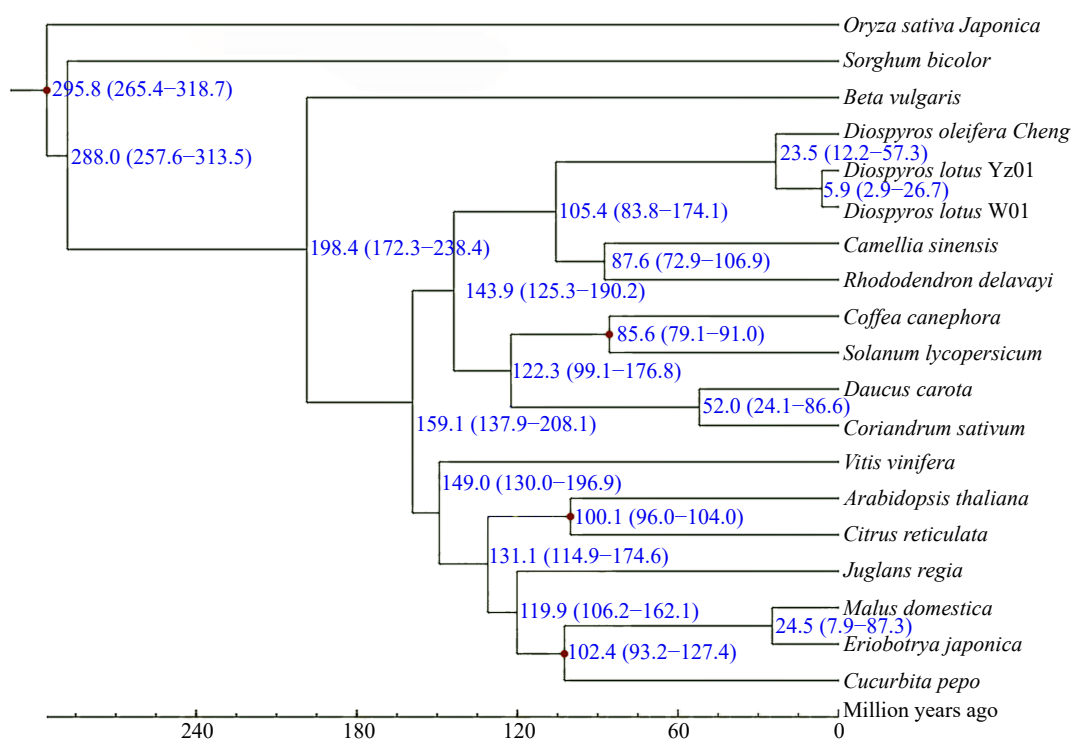


Fig. 6 Phylogenetic tree of seedless and seeded *Diospyros lotus* and 17 other species constructed using the maximum-likelihood method. Estimated species divergence times (million years ago) and 95% confidence intervals are labeled at each branch site. Blue numbers on branches indicate the estimated divergence times. Red dots indicate the divergence times estimated based on fossil evidence.

monoterpenoid biosynthesis, glucosinolate biosynthesis and alpha-linolenic acid metabolism were among the enriched KEGG pathways (Supplemental Table S16). Moreover, the most enriched GO terms were oxidation-reduction process, ionotropic glutamate receptor activity, extracellular ligand-gated ion channel activity and glutamate receptor activity (Supplemental Table S17). Several KEGG pathways were enriched among the 1,951 gene families predicted to have contracted in seeded, including one carbon pool by folate, nitrogen metabolism and anthocyanin biosynthesis (Supplemental Table S18), whereas the enriched GO terms were catalytic activity, serine-type endopeptidase activity and copper ion binding (Supplemental Table S19).

Seedlessness may be associated with genes related to pollen and pollination, fertilization and various hormone regulators. The expanded and contracted seedless *D. lotus* gene families included those associated with fatty acid elongation and the MAPK signaling pathway, which are important for regulating plant hormones. Lipids are part of hormone precursors, whereas auxin, ethylene and abscisic acid are correlated with the MAPK signaling pathway^[48,49]. Therefore, analyses of the genes enriched in these pathways may provide new insights into the formation of seedless fruits.

Database construction

Diospyros, which is the largest genus in the family Ebenaceae, comprises more than 500 economically valuable species widely distributed in the tropics and subtropics, including approximately 300 species in the Asia-Pacific region, 98 species in Madagascar and the Comoros, 94 species in mainland Africa, about 100 species in the Americas, 15

species in Australia, and 31 species in New Caledonia^[50,51]. Many persimmon species have been studied, but relatively little research has focused on the genome. The recent increase in genome resources has produced a wealth of data for in-depth analyses of the biology and evolution of *Diospyros* plants, but obtaining and using these resources remains difficult. Thus, we developed the *Diospyros* Genome Database (<http://www.persimmongenome.cn>) as the first comprehensive database for *Diospyros* plant genomes. This database provides tools for browsing genomes (JBrowse), searching sequence databases (BLAST), and designing primers. To better serve the research community, we continue to update our database and develop new tools (Fig. 7).

CONCLUSIONS

We applied Illumina and PacBio sequencing platforms and Hi-C technology to assemble chromosome-level reference genomes for seeded and seedless *D. lotus*. The resulting seedless and seeded *D. lotus* genomes comprised 617.66 and 647.31 Mb, respectively. The assembled seeded genome included 23,193 protein-coding genes, 219 miRNAs, 826 tRNAs, 2,386 rRNAs, 1,371 snRNAs, 424 expanded gene families, and 1,951 contracted gene families. The assembled seedless genome included 21,684 protein-coding genes, 146 miRNAs, 496 tRNAs, 719 rRNAs, 792 snRNAs, 490 expanded gene families, and 1,497 contracted gene families. We predicted that *D. lotus* and *D. oleifera* diverged approximately 23.5 million years ago, whereas seeded and seedless *D. lotus* diverged about 5.9 million years ago. The high-quality *D. lotus* genomes assembled in this study will be useful for future

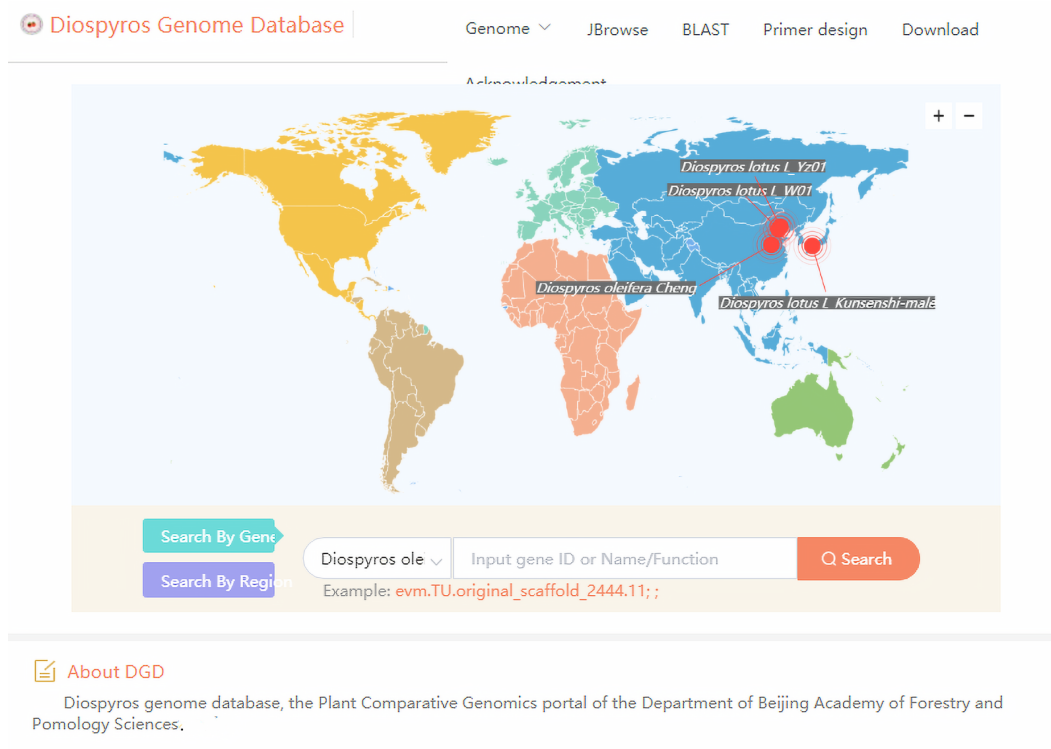


Fig. 7 The user interface of *Diospyros* Genome Database for browsing genomes, searching for homologous sequences and designing primers.

research on important agronomic traits among *Diospyros* species. Furthermore, comparisons between seeded and seedless *D. lotus* genomes will enable researchers to clarify the mechanisms underlying seedlessness.

ACKNOWLEDGMENTS

This work was financially supported by the Presidential Foundation of Beijing Academy of Forestry and Pomology Sciences (grant no. LGY201901) and the Special Fund for the Construction of Scientific and Technological Innovation Capability (KJCX20200114).

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Information accompanies this paper at (<http://www.maxapress.com/article/doi/10.48130/FR-2021-0009>)

Dates

Received 25 March 2021; Accepted 14 May 2021; Published online 27 May 2021

REFERENCES

1. Christophel D. 1982. Earliest floral evidence for the Ebenaceae in Australia. *Nature* 296:439–41
2. Duangjai S, Wallnöfer B, Samuel R, Munzinger J, Chase MW. 2006. Generic delimitation and relationships in Ebenaceae sensu lato: evidence from six plastid DNA regions. *American Journal of Botany* 93:1808–27
3. Turner B, Munzinger J, Duangjai S, Temsch EM, Stockenhuber R, et al. 2013. Molecular phylogenetics of New Caledonian *Diospyros* (Ebenaceae) using plastid and nuclear markers. *Molecular Phylogenetics and Evolution* 69:740–63
4. Loizzo MR, Said A, Tundis R, Hawas UW, Rashed K, et al. 2009. Antioxidant and Antiproliferative Activity of *Diospyros lotus* L. Extract and Isolated Compounds. *Plant Foods Hum. Nutr.* 64:264
5. Rauf A, Uddin G, Siddiqui BS, Muhammad N, Khan H. 2014. Antipyretic and antinociceptive activity of *Diospyros lotus* L. in animals. *Asian Pac. J. Trop. Biomed.* 4:S382–S386
6. Yang Y, Yang T, Jing Z. 2015. Genetic diversity and taxonomic studies of date plum (*Diospyros lotus* L.) using morphological traits and SCoT markers. *Biochem. Syst. Ecol.* 61:253–59
7. Cho BO, Yin HH, Park SH, Byun EB, Ha HY, et al. 2016. Anti-inflammatory activity of myricetin from *Diospyros lotus* through suppression of NF- κ B and STAT1 activation and Nrf2-mediated HO-1 induction in lipopolysaccharide-stimulated RAW264.7 macrophages. *Biosci. Biotechnol. Biochem.* 80:1520–30
8. Zhou R, Zhang X, Hu H, Li G, Song R. 2016. Plant regeneration from leaves of seedless date plum (*Diospyros lotus* L.). *Northern Horticulture* 40(22):104–6
9. Ali S, Khan AS, Raza SA, Naveed R, Rehman R. 2013. Innovative breeding methods to develop seedless citrus cultivars. *International Journal of Biosciences* 3:191–201
10. Mesejo C, Martínez-Fuentes A, Reig C, Rivas F, Agustí M. 2006. The inhibitory effect of CuSO₄ on *Citrus* pollen germination and pollen tube growth and its application for the production of seedless fruit. *Plant Science* 170:37–43
11. Sugiyama K, Morishita M. 2000. Production of seedless watermelon using soft-X-irradiated pollen. *Scientia Horticulturae* 84:255–64

12. Mesejo C, Reig C, Martínez-Fuentes A, Agustí M. 2010. Parthenocarpic fruit production in loquat (*Eriobotrya japonica* Lindl.) by using gibberellic acid. *Scientia Horticulturae* 126:37–41
13. Doyle JJ, Doyle JL. 1986. A rapid DNA isolation procedure for small quantities of fresh leaf tissues. *Phytochemical Bulletin* 19:11–15
14. Koren S, Walenz PB, Berlin K, Miller JR, Bergman NH, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research* 27:722–36
15. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10:563–69
16. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963
17. Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19:460
18. Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27:764–70
19. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–59
20. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, et al. 2015. HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Research* 4:1310
21. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356:92–95
22. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754–60
23. Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. *BMC Bioinformatics* 13:238
24. Simão F, Waterhouse R, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–12
25. Ou S, Jiang N. 2017. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* 176:1410–22
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–79
27. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, et al. 2012. Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology* 30:78–82
28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. 2010. The genome analysis toolkit: A mapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297–303
29. Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* 25:4.10.1–4.10.14
30. Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27:573–80
31. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, et al. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research* 34:W435–W439
32. Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF. 2006. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology* 4:41
33. Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–11
34. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. 2012. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and cufflinks. *Nature Protocols* 7:562–78
35. Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics* 48:4.11.1–4.11.39
36. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25:955–64
37. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, et al. 2018. Non-coding RNA analysis using the Rfam database. *Current Protocols in Bioinformatics* 62:e51
38. Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25:1335–37
39. Wang Y, Tang H, DeBarry J, Tan X, Li J, et al. 2012. MScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* 40:e49
40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10
41. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157
42. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–97
43. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307–21
44. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24:1586–91
45. De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269–71
46. Martin G, Carreel F, Coriton O, Hervouet C, Cardic C, et al. 2017. Evolution of the banana genome (*Musa acuminata*) is impacted by large chromosomal translocations. *Molecular Biology and Evolution* 34:2140–52
47. Copley RR, Letunic I, Bork P. 2002. Genome and protein evolution in eukaryotes. *Curr. Opin. Chem. Biol.* 6:39–45
48. Danquah A, de Zelicourt A, Colcombet J, Hirt H. 2014. The role of ABA and MAPK signaling pathways in plant abiotic stress responses. *Biotechnology Advances* 32:40–52
49. Roudier F, Gissot L, Beaudoin F, Haslam R, Michaelson L, et al. 2010. Very-long-chain fatty acids are involved in polar auxin transport and developmental patterning in *Arabidopsis*. *The Plant Cell* 22:364–75
50. Duangjai S, Samuel R, Munzinger J, Forest F, Wallnöfer B, et al. 2009. A multi-locus plastid phylogenetic analysis of the pantropical genus *Diospyros* (Ebenaceae), with an emphasis on the radiation and biogeographic origins of the New Caledonian endemic species. *Mol. Phylogenet. Evol.* 52:602–20
51. Rauf A, Uddin G, Patel S, Khan A, Halim SA, et al. 2017. *Diospyros*, an under-utilized, multi-purpose plant genus: A review. *Biomedicine Pharmacotherapy* 91:714–30



Copyright: © 2021 by the author(s). Published by Maximum Academic Press, Fayetteville, GA.

This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.