Chromosome Level Genome Assembly Reveals *Cassia Tora* May Be an Ancient Species in Leguminosae

Caicai Lin
Shandong Agricultural University
Piyi Xing
Shandong Agricultural University
Changhao Zhou
Shandong Agricultural University
Xingfeng Li
Shandong Agricultural University
Zhenqiao Song (szq@sdau.edu.cn)
Shandong Agricultural University

Research article

Keywords: Cassia tora, genome, Hi-C, PacBio sequencing, Pseudochromosomes

Posted Date: December 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-134025/v1

Abstract

Background: Cassia tora L. is an annual leguminous plant. Its seeds had wide utility in herbal medicine in Asian, but is usually regard as a potent, invasive weed which even helps farmers eliminate other parasitic weed species. However, compared with other important legume crops, *C. tora* is far from fully developed and the genetic basis is greatly lacking. A reference genome sequence will be great valuable resource for its genome evolution, genetic breeding and development.

Results: Here, we *de novo* assembled a chromosome-scale genome for *C. tora* by combining PacBio sequencing technology with chromatin interaction mapping, resulting in 621-Mb genome with a contig N50 of 2.5 Mb, of which 77.44% was ordered and oriented on 13 pseudochromosomes. The genome contained 32,361 protein-coding genes with a repetitive DNA content of approximately 58.15%. *Gypsy*-type LTRs constituted the largest subfamily. LTR insertion events seldom occurred in this genome over the past 10 million years. Comparative genomic analyses showed that *C. tora* diverged ~95 million years ago (MYA), which revealed it has the most distant genetic relationship with 11 other legumes. Compared with other legume crops, *Cassia tora* is maybe an ancient species in Leguminosae.

Conclusions: The high-quality reference genome sequence reported here furnishes unprecedented insights into genome dynamics and provides an important basis for future research on legume genome evolution.

Background

Cassia tora L. (syn. *Senna tora* Roxb.) is an annual leguminous weed in the subfamily Caesalpinioideae [1]. It was originally introduced from the tropical Americas and is now extensively distributed throughout China, Korea, India, Southeast Asia, and the Southwest Pacific [2]. *C. tora* grows easily and rapidly, is very stress tolerant, and thrives in dry soil from elevations of sea level to 1,800 m. It interferes with the growth of agriculturally important crops and economically useful plants and depletes soil fertility [3]. *C. tora* disperses widely and occupies substantial tracts of farmland that could otherwise be allocated towards the cultivation of various types of valuable vegetation.

However, *C. tora* may actually be beneficial to farmers in certain ways. As it starts to grow, *C. tora* occupies all adjacent spaces and prevents other weeds from establishing there. Natural green grass seldom grows in areas overtaken by *C. tora*. Thus, it naturally removes and replaces weeds. *Parthenium hysterophorus* has disastrous effects on agricultural fields and may trigger skin and respiratory ailments. Though it has been difficult to eradicate, *P. hysterophorus* is effectively inhibited by *C. tora*. Aqueous extracts of *C. tora* may contain allelochemicals that suppress the germination of seeds produced by other varieties of weeds[4]. Moreover, the application of *C. tora* seed gum as a natural coagulant of raw and undiluted pulp and paper mill effluent (PPME) has been investigated [5].

C. tora also contains anthraquinones and has been used in traditional Chinese and Ayurvedic medicine. The dry and ripe seeds of *C. tora* are known as ‘Juemingzi’ and considered traditional Chinese medicine

(TCM). Their use was recorded in ‘Shennong Ben Cao Jing’ and they have been administered to treat headache, dizziness, dysentery, and eye disease for > 2,000 y. Modern pharmacology has demonstrated that Juemingzi extracts have antioxidant, hypolipidemic, neuroprotective, and other physiological efficacy [6]. Thus, Juemingzi has become a very popular TCM for the treatment of mutagenicity, genotoxicity, hepatotoxicity, and acute inflammatory diseases [7–10]. Various pharmacologically active chemical constituents have been detected and identified in *C. tora* seeds including anthraquinones, terpenoids, flavonoids, lipids, phenolic compounds, amino acids, and polysaccharides [11, 12]. *C. tora* seeds are also used as an ingredient in functional and popular herbal beverages in Asia. Purified *C. tora* endosperm flour has been used as a food thickener (E-427) [13].

As a widely distributed legume plant, compared with other legumes, its use is far from being developed, the genetic basis is greatly lacking, and the breeding of varieties has not been reported. The genome evolution of leguminous plants has aroused widespread concern [14–19], but the genome evolution of *C. tora* is still unclear.

High-quality genomes are the foundations of genetic and genome-wide studies [20–23]. Here, we assembled a high-quality, chromosome-scale genome sequence for *C. tora* using a combination of single-molecule real-time (SMRT), and Hi-C sequencing. This genome sequence will facilitate genomic research, metabolic engineering, and cultivar improvement in *C. tora*.

Results

Genome size estimate

We selected a landrace line JMZ1 for genome size estimate. The highest peak provided a peak depth of 74 for genome size estimation. As the total number of k-mers was 45,035,768,647, the genome size was calculated to be ~ 601.44 Mb. Therefore, the sequenced Illumina reads (56.18 Gb) provided ~ 93 × coverage. The estimated repetitive sequence content was ~ 58.61% based on a peak depth of 142. As no heterozygosity peak was found, the estimated heterozygosity was ~ 0.02% based on a half-depth of 74. The GC content of the genome of this species was ~ 37.30% (Supplementary Fig. S1; Supplementary Table S1;).

Genome sequencing and assembly

De novo DSS3 genome assembly was performed using a combination of PacBio single-molecule, real-time (SMRT) sequencing and chromatin interaction mapping (Hi-C). We obtained 56.40 Gb raw data by PacBio SMRT sequencing (Supplementary Table S2). After filtering out low-quality data, we generated 55.23 Gb of high-quality subreads with an average contig length of 8.85 kb and N50 = 13.42 kb. This large dataset covered ~ 93-fold of the predicted genome size. We optimized the genome with Quickmerge and generated a preliminary genome assembly capturing 621.06 Mb with contig N50 = 4.16 Mb and 4,320 contigs (Supplementary Table S3).

A total of 119 Mb of read pairs was collected via Illumina sequencing and 35.94 Gb of clean Hi-C reads were generated with 57-fold coverage. The scaffolds were broken into 50-kb fragments of equal length and reassembled by Hi-C data analysis. The low- and middle Hi-C coverage depths in this region were localized and identified as the error points. The aforementioned genome assembly was corrected and a final genome was generated that captured 621.06 Mb with contig N50 = 2.5 Mb, 4,473 contigs, and scaffold N50 = 29.35 Mb (Table 1).

Table 1
C. tora genome assembly information.

Chromosome number (n)	**13**
Genome size (Mb)	568.59
Contig number	4473
Contig N50 (kb)	**2,500**
Contig N90 (kb)	43
Contig max (kb)	**17,300**
Scaffold number	4,229
Scaffold N50 (kb)	**29,349**
Scaffold N90 (bp)	43,629
Scaffold max (bp)	44,760,356
GC content (%)	35.96
Note: Scaffolds, including unplaced contigs, are defined as input contigs not placed by the optical map.	

The Hi-C data reads mapping to the assembly genome consisted of 212 Mb pairs corresponding to 88.38% of the genome pairs. There were 63 Mb of unique mapped read pairs accounting for 52.90% and meeting the requirements of the subsequent analysis (Supplementary Table S4). For the unique mapped read pairs, we evaluated the library data, deleted invalid dangling end-, religation-, self-cycle-, and dumped pairs, and identified 52.85 Mb of valid interaction pairs to anchor the aforementioned genome sequences to pseudochromosomes. A total of 568.59 Mb of sequences were mapped to 13 pseudochromosomes accounting for 91.55% of the whole-genome sequence. Of these, 440.34 Mb was ordered and oriented with pseudochromosomes in the size range of 26.33–55.29 Mb (Table 2). The Hi-C assembly heatmap of the chromosomes displayed 13 distinct groups which denoted high-quality assembly (Fig. 1).

Table 2
Clustered sequence data of each pseudochromosome.

Group	Sequence Number	Sequence Length (bp)
Lachesis Group 0	281	57,814,068
Lachesis Group 1	138	53,811,674
Lachesis Group 2	231	55,297,032
Lachesis Group 3	234	51,878,843
Lachesis Group 4	161	45,427,400
Lachesis Group 5	294	48,951,038
Lachesis Group 6	144	39,736,525
Lachesis Group 7	191	41,298,315
Lachesis Group 8	137	35,748,794
Lachesis Group 9	135	35,033,854
Lachesis Group 10	227	38,596,663
Lachesis Group 11	61	26,337,058
Lachesis Group 12	71	38,666,604
Total Sequences Clustered	2,305	568,597,868
Total Sequences Ordered and Oriented	257	440,341,984

Note: Ratio% of first two items shows % of clustered sequence number or length compared to contig number or genome size. Ratio% of next two items shows % of ordered and oriented sequence number or length compared to cluster sequence number or sequence.

We used the raw Illumina HiSeq paired-end reads to validate the genome assembly. Of these, 85% could be properly mapped to the assembly (Supplementary Table S5). Thus, the genome assembly contained comprehensive genome information. The Core Eukaryotic Genes Mapping Approach (CEGMA) and Benchmarking Universal Single-Copy Orthologs (BUSCO) methods were used to assess genome assembly accuracy and completeness. The CEGMA v. 2.5 alignment showed that 448 CEG proteins (97.82%) with high-confidence hits and 229 (92.34%) highly conserved full-length sequences were contained in the genome assembly. A BUSCO analysis revealed that we obtained > 93.89% (1,352/1,440 BUSCOs) genome coverage.

Annotation of the C. tora *genome sequence*

Using homology-based and *de novo* approaches, we identified 361.17 Mb of repetitive elements accounting for 58.15% of the genome. The main repetitive sequences were transposable elements (TEs) of which the most abundant were long terminal repeats (LTRs) (Supplementary Table S6). Among the Class I elements (retrotransposons), the *Gypsy*-, LARD, and *Copia*-type LTRs represented the three largest subfamilies (19.88%, 7.12%, and 6.81%, respectively). Among the Class II elements (DNA transposons), the CACTA repeats and Helitron-type elements occupied 1.68% and 1.39% of the genome, respectively. The remaining genome consisted of putative host genes (1.74%) and simple sequence repeats (SSRs; 6.82%). Most of these TEs were distributed in the central parts of the chromosomes. Most of the repeat sequence types were unequally distributed along the chromosomes. Their densities were higher in the centromeric than the telomeric regions on all pseudomolecules with similar patterns (Fig. 2). Overall, the centromeric regions were enriched in various repeat elements.

We identified 11.64% unclassified repeats in the genome. Thus, many new LTR types could be discovered in plants. We hypothesize that these TEs are different from those of the ancestral legume species and may have contributed to diversification and speciation in these plants.

A combination of homology-based, *de novo*, and RNA-seq approaches predicted 32,361 protein-coding genes that covered 20.37% of the genomic sequence length (126.52 Mb) (Supplementary Table S7). In total, 87.76% of the predicted genes were supported by homology-based- and RNA-seq data derived from a mixture of leaf, stem, flower, and pod tissues (Supplementary Fig. S2). The average gene size and exon number per gene were 3,875 bp and 5.09, respectively (Supplementary Table S8). The average exon and intron lengths were 239.09 bp and 528.87 bp, respectively. Of the predicted genes, 29,163 (90.12%) were functionally annotated in the NR, TrEMBL, GO, KOG, and KEGG databases (Supplementary Table S9). Most of the protein-coding genes were concentrated at the distal chromosome regions and were relatively sparse in the proximal regions on pseudochromosomes with similar distribution patterns (Fig. 2). Thus, the gene distribution was non-random and resembled that for other plant species reported [23, 24].

We predicted 3,948 non-coding genes including 632 transfer RNAs, 446 ribosomal RNAs, 88 microRNAs, and 2,782 pseudogenes with evidence of transcription but no consistent coding sequence. We also detected 2,980 motifs and 34,761 domains (Supplementary Table S10).

Comparative analysis of the genomes of C. tora *and other plant species*

A gene family cluster analysis of the complete gene sets was performed on *C. tora* and certain sequenced legume species such as *Arachis duranensis*, *Arachis ipaensis*, *Cajanus cajan*, *Cicer arietinum*, *Glycine max*, *Glycyrrhiza uralensis*, *Lotus japonicus*, *Medicago truncatula*, *Phaseolus vulgaris*, *Vigna angularis, Vigna radiata*, and the non-legume *Vitis vinifera.*

The 26,757 predicted genes were classified into 15,619 gene families with unique gene family number of 780 to *C. tora* (Supplementary Table S11). Four hundred and six gene families expanded and 5,363 gene families contracted in *C. tora*. The expanded gene families were enriched mainly in the GO (gene ontology) categories ‘metabolic process’, ‘cellular process in biological process’, ‘membrane in cellular

component', and 'catalytic activity and binding in molecular function' (Supplementary Fig. S3). The contracted gene families in *C. tora* were enriched in 'biological death process', 'antioxidant activity', and 'nutrient reservoir activity'.

A phylogenetic tree was constructed based on the single-copy orthologous genes shared by *C. tora* and 11 other legumes (Fig. 3). The legumes differentiated from grapes over 116 MYA. *Cassia tora* differentiated from 11 other legumes ~ 95 MYA. Therefore, *C. tora* is only distantly related to other legumes. After ~ 20 MYA, peanut differentiated from the other nine legumes. The latter then evolved into two clusters. Five vegetable legume species grouped into one cluster and differentiated only ~ 33–4.7 MYA. The other cluster included four legumes (alfalfa for leaf use, licorice for root use, lobelia for leaf use, and chickpea for bean use) that differentiated as early as 56–34 MYA.

LTR family expansion number and time were analyzed for the genomes of *C. tora* and 11 other plant species (Fig. 4). Using *C. cajan* as a reference, no distinct peak LTR amplification activity was observed in *C. tora* over 10 MYA. Thus, the *C. tora* genome was stable and inactive compared to the other 11 species.

Collinearity analysis between C. tora and A. ipaensis

Cassia tora and *A. ipaensis* (subfamily Faboideae) diverged ~ 95 MYA. A synteny analysis between *C. tora* and *A. ipaensis* revealed that 33.79% of their genes were collinear in 490 blocks (Fig. 5). The genes on one peanut chromosome were dispersed over four *C. tora* chromosomes. This finding reflected poor collinearity and demonstrated substantial differences between genomes.

The 4DTv (fourfold degenerate transversion) distribution of the homologous gene pairs within these syntenic blocks suggested that *C. tora* has not undergone any recent lineage-specific whole-genome duplication (WGD) event. Rather, it experienced the paleohexaploidy event (γ) common to all eudicots (Fig. 6).

Discussion

Compared with other legumes, the genetic research of *C. tora* is still in its infancy. Although its nutritive and pharmacological properties of the active metabolites have been investigated [7–10] and their biosynthetic pathways have been preliminarily characterized [25], to the best of our knowledge, the *C. tora* genetic basic research is very limited.

Single-molecule sequencing technology combined with Hi-C assembly technology [26, 27] has been widely used in the construction of a chromosome level high-quality genome [28–31]. A high quality genome at chromosome level lays a very important foundation for the further research and development of a species [19, 24, 32, 33]. With the dividend of the decreasing cost of sequencing, we were able to sequence the species with Single-molecule sequencing technology and assemble it with Hi-C technology to form a high-quality genome. The ancestral legume karyotype was predicted consisting of a minimum

of 19 proto-chromosomes [19]. In the study, the Hi-C assembly displayed 13 distinct groups. We propose the *C. tora* genome maybe have been massively rearranged during evolution.

Genome expansion in plants is primarily driven by polyploidization (whole-genome duplication events) and the proliferation of TEs [34]. Based on the single-copy orthologous genes shared by *C. tora* and 11 other legumes (Fig. 3), these legume species differentiated from grapes on about 116 MYA. *C. tora* differentiated from 11 other legumes ~ 95 MYA which revealed *C. tora* has the most distant relationship with other legumes. After ~ 20 MYA, peanut differentiated from the other nine legumes. Other legume species differentiated in 75–80 MY, which was consistent with recent report about pea genome [19]. Among the other 11 legumes, *A. ipaensis* has the closest relationship with it. Analysis of genome sequences of different legumes indicated that whole genome duplication has played an important role in evolution of legume genome [17]. The 4DTv distribution of the homologous gene pairs within these syntenic blocks suggested that *C. tora* has not undergone any recent lineage-specific whole-genome duplication (WGD) event except for the paleohexaploidy event (γ) common to all eudicots. A synteny analysis between *C. tora* and *A. ipaensis* reflected poor collinearity.

Repetitive elements were major drivers in the evolution of these large genomes such as Pea genome (~ 4.45 Gb) [19]. In the Fabeae tribe of Leguminosae, genome dynamics are dominated by a single lineage of Gypsy elements that account for 57% of the variation in genome size in this clade [34, 35]. In *C. tora* genome, Fifty-eight percent of the genomic sequences belongs to repetitive. The main repetitive sequences were TEs, of which the most abundant were long terminal repeats (LTRs) and the Gyps -type LTRs represented the three largest subfamilies, which was consistent with previous reports [34, 35]. Compared to the other 11 species, LTR amplification activity peak in *C. tora* was not observed in over 10 MYA, which reflected a feature of genome formation which is not active. Another possible explanation is that the LTR-TEs in *C. tora* had undergone lower artificial selection than those in the other species. Moreover, as *C. tora* naturally self-pollinates, it has low heterozygosity (0.02%). The foregoing genomic analysis indicated that the *Cassia tora* genome is ancient and stable.

In addition to comparative genomic research, this genome also will provide an important foundation for the later research, such as characterizing and amplifing the genes encoding the anthraquinones in C. tora, or identifing the genes responsible for its virulence against other plants (beneficial or detrimental).

Conclusion

Here, we produced a high-quality genome sequence for *C. tora* that may serve as a vital tool to analysis features of the *C. tora* genome and help to characterize legume family and establish deep syntenic relationships among different taxa of legumes. The foregoing analysis indicated that the *C. tora* was a relatively stable and relatively old genome. The valuable genome sequence will also facilitate the characterization of its many excellent traits, enhance genetic improvement and allow more efficient use of the genus.

Methods

Materials

A landrance line JMZ1 was originated from Tai Mountain and was identified as *Cassia tora* by Dr. Jianhua Wang, then collected and planted in medicinal plant garden of Shandong Agricultural University in Shandong Province, China. We have acquired permission to collect plant samples of JMZ1 from Dr. Jianhua Wang. Seeds of the landrace line were harvested and deposited in Germplasm Bank of Medicinal Botanical Garden in Shandong Agricultural University (Deposition number SDAUMD101). JMZ1 was selected for genome and Hi-C sequencing. RNA sequencing for the prediction of gene structures was carried out in mixed DNA samples with the same amount of DNA from four tissues, including the roots, stem, leaves and flowers.

Illumina sequencing and estimation of C. tora genome size

Genomic DNA was extracted from the leaf tissues of *C. tora* by the cetyltrimethylammonium bromide (CTAB) method. The experiments were performed according to the standard DNA library preparation and sequencing protocols provided by Illumina (San Diego, CA, USA). Two 270-bp libraries were constructed using genomic DNA from JMZ1 (Supplementary Table 1a). All of the ~56.18 Gb of clean reads obtained from the Illumina platforms were subjected to 19-mer frequency distribution analysis. The K-mer distribution can be used to infer the estimated genome size using the following formula: genome size = k-mer number/peak depth.

Genome sequencing, assembly, and scaffolding

The libraries for SMRT sequencing were constructed using DNA isolated from leaf tissue according to the standard protocol provided by Pacific Biosciences (Menlo Park, CA, USA). This protocol comprised library construction and sequencing as well as library and sample quality detection.

De novo assembly of the PacBio reads was initially conducted with the Canu pipeline (v1.5) and Falcon (https://github.com/ruanjue/wtdbg) assemblers. Their parameters were set for high consistency and continuity. The sequence assembled by Canu was then used to assemble the genome in Wtbdg (v1.2.8) (https://github.com/ruanjue/wtdbg). Illumina paired-end reads were then used in assembly polishing with Pilon v. 1.17 to correct the assembly [36]. Statistical information is shown in Supplementary Table 2.

Hi-C sequencing, assembly, scaffolding, and evaluation

Hi-C experiments were performed as previously reported [37]. Both valid and non-valid interaction pairs were identified by HiC-Pro [27]. Hi-C data were aligned with the initially assembled genome (v. 1.1) by BWA v. 0.7.10-r789 to obtain mapped data [38]. The paired-end reads were mapped onto v. 1.2 PacBio contigs which were then grouped into13 chromosome clusters and scaffolded using Lachesis software [37].

Genome quality evaluation

Illumina reads were used to map and confirm deciphered sequence coverage (Supplementary Table 5). The completeness of the final genome assembly was evaluated by CEGMA v.2.3 [32]and Benchmarking Universal Single-Copy Orthologs (BUSCO v2.0) [39].

Transposable element annotation

Homology-based- and *de novo* strategies were applied to identify repetitive sequences in the genome. *De novo* prediction programs including LTR-Finder (v1.0.2) [40], MITE [41], RepeatScout (v1.0.5) [42], and PILER (v1.0) [43] were used. The repetitive sequences were segregated into various categories with PASTEClassifier [44].

Genetic model prediction, evaluation, and annotation

Homology-based, *de novo*, and RNA-seq genetic model prediction methods were used in the present study. For the homology-based gene prediction model, geMoMa v1.3.1 [45] to predict gene structures. Genscan [46], Augustus v2.4 [47], GlimmerHMM v3.0.4 [48], GeneID v1.4 [49], and SNAP [50] were used for *de novo* prediction, for the transcriptome-based prediction with PASA v. 2.0.2 [45].

All of the predicted gene structures were integrated into a consensus set with EVidenceModeler [51]. Finally, 32,361 genes were obtained and annotated with BLAST (E value $\leq 1 \times 10^{-5}$) according to their homology alignments against the annotation databases NR [52], KOG [53], KEGG [54], GO [55]. InterProScan v5.8-49.0 [56] was used to predict motif annotation information by alignment with databases such as the PROSITE [57], HAMAP [58], Pfam [59], PRINTS [60], ProDom [61], SMART [62], TIGRFAM [63], PIRSF [64], SUPERFAMILY [65], CATH-Gene3D [66].

Comparative genome analysis

OrthoMCL (v2.0) [67] was used to analyze the family classification of protein sequences. PHYML [68] software was used to construct the phylogenetic tree. CAFE (v4.2) [69] was used to analyze gene family contraction and expansion. LTR_FINDER (v1.0.2) [40] were used to search the LTR sequences. The molecular clock was selected as 7.3×10^{-9}. The 4DTv distances of the orthologous gene pairs in *C. tora* and *A. ipaensis* were calculated with the HKY (Hasegawa, Kishino, and Yano) substitution model based on the codon alignment.

Abbreviations

MYA: Million years ago

TCM: traditional Chinese medicine

SMRT: single-molecule real-time

Hi-C: chromatin interaction mapping

CEGMA: Core Eukaryotic Genes Mapping Approach

BUSCO: Benchmarking Universal Single-Copy Orthologs

TEs: transposable elements

LTRs: long terminal repeats

KEGG: Kyoto Encyclopedia of Genes and Genomes

FPKM: Fragments per kilobase of transcript per million fragments

GO: Gene ontology

CC: Cellular component

MF: Molecular function

BP: Biological process

4DTv: four-fold degenerate transversion

WGD: whole-genome duplication

Declarations

Acknowledgments

Not applicable.

Funding

This work was supported by the National Natural Science Foundation of China (No. 81872949), the National Natural Science Foundation of Shandong Province (No. 2019HM081), and the Double First-Class Construction Project of Shandong Agricultural University to Z.S. and X.L.

Availability of data and materials

The assembly and annotation data of *C. tora* in the GenomeWarehouse in BIG Data Center under Project numbers PRJCA003150, which are accessible at https://bigd.big.ac.cn/gwh.

Ethics approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author contributions

Z.S. and X.L. conceived and designed the research, coordinated the study, and wrote the manuscript. Z.S. and C.L. performed the long-read sequencing and genome assembly. C.L., P.X. and Z.S. designed the Hi-C experiments, produced assembly scaffolds from the data, and polished the final reference assembly. Z.S. and P.X. performed the genome annotation and evaluation analysis and assisted with downstream analysis. C.L. sowed plant samples for sequencing. C.Z. assisted with the generation of additional files for the manuscript and contributed to data analysis. All authors read and approved the final manuscript.

Declaration

This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal.

References

1. Tripathi VR, Kumar S, Garg SK. A study on trypsin, Aspergillus flavus and Bacillus sp. protease inhibitory activity in *Cassia tora* (L.) syn *Senna tora* (L.) Roxb. seed extract. BMC Complement Altern Med. 2011;11:56.
2. Vadivel W, Kunyanga CN, Biesalski HK. Antioxidant potential and type II diabetes-related enzyme inhibition of *Cassia tora*: effect of indigenous processing methods. Food Bioprocess Technol. 2011; 5: 2687-96.
3. Cock MJW, Evans HC. Possibilities for biological control of *Cassia tora* and *obtusifolia*. Tropical Pest Management. 1984;30(4):339-50.
4. Bhatt PS, Singh B, Todaria NP. Effects of *Cassia tora* on the germination and growth of *Parthenium hysterophorus* Allelopathy Journal. 2006;17(2):303-10.
5. Subramonian W, Wu TY, Chai SP. A comprehensive study on coagulant performance and floc characterization of natural *Cassia tora* seed gum in treatment of raw pulp and paper mill effluent. Ind Crop Prod. 2014;61: 317-24.
6. Kim SJ, Kim KW, Kim DS, Kim MC, Jeon YD, Kim SG, et al. The protective effect of *Cassia obtusifolia* on DSS-induced colitis. The American Journal of Chinese Medicine. 2011; 39(3):565-77.
7. Jung HA, Ali MY, Jung HJ, Jeong HO, Chung, HY, Choi, JS. Inhibitory activities of major anthraquinones and other constituents from *Cassia tora* against β-secretase and cholinesterases. Journal of Ethnopharmacology. 2016; 191:152-60.

8. Jung HA, Ali MY, Choi J. Promising Inhibitory Effects of Anthraquinones, Naphthopyrone, and Naphthalene Glycosides from *Cassia tora* on α-Glucosidase and Human Protein Tyrosine Phosphatases 1B. Molecules. 2016; 22 (1): 28.
9. Yi JH, Park HJ, Lee S, Jung JW, Kim B, Lee YC, et al. *Cassia tora* seed ameliorates amyloid β-induced synaptic dysfunction through anti-inflammatory and Akt/GSK-3β pathways. J Ethnopharmacol. 2016; 178: 50-57.
10. Paudel P, Jung HA, Choi JS. Anthraquinone and naphthopyrone glycosides from *Cassia obtusifolia* seeds mediate hepatoprotection via nrf2-mediated ho-1 activation and mapk modulation. Archives of Pharmacal Research. 2018;41(6):677-89.
11. Sob SVT, Wabo HK, Tchinda AT, Tane P, Ngadjui BT, Ye Y. Anthraquinones, sterols, triterpenoids and xanthones from *Cassia tora*. Biochem Syst Ecol. 2010;38:342-45.
12. Feng L, Yin J, Nie S, Wan Y, Xie M. Fractionation, physicochemical property and immunological activity of polysaccharides from *Cassia tora*. International Journal of Biological Macromolecules.2016; 91: 946-53.
13. Huang YL, Chow CJ, Tsai YH. Composition, characteristics, and in-vitro physiological effects of the water-soluble polysaccharides from Cassia seed. Food Chem. 2012;134:1967-72.
14. Lavin M, Herendeen PS, Wojciechowski MF. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. Syst Biol. 2005, 54:575-94.
15. Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, et al. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* PNAS.2016; 103(40):14959-64.
16. Goffard N, Weiller G. Functional analysis of legume genome arrays. Methods Mol Biol. 2013; 1069:59-66.
17. Kumar J, Srivastava E, Singh M, Pratap A. Genomics in Studying the Legume Genome Evolution. In: Gupta S., Nadarajan N., Gupta D. (eds) Legumes in the Omic Era. Springer New York; 2013.287-300
18. Ksikiewicz M, Zielezinski A, Wyrwa K, Szczepaniak A, Rychel S, Karlowski W, et al. Remnants of the legume ancestral genome preserved in gene-rich regions: insights from *Lupinus angustifolius* physical, genetic, and comparative mapping. Plant Mol Biol Report. 2015; 33(1):84-101.
19. Kreplak J, Madoui MA, Cápal P, Novák P, Labadie K, Aubert G, et al. A reference genome for pea provides insight into legume genome evolution. Nat Genet. 2019; 51:1411-22.
20. Wang L, Yu S, Tong C, Tong CB, Zhao YZ, Liu Y, et al. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. Genome Biology.2014; 15(2): R39.
21. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics. 2017; 49(7):1099-1106.
22. Yuan Z, Fang Y, Zhang T, Fei Z, Han F, Liu C, et al. The pomegranate (*Punica granatum*) genome provides insights into fruit quality and ovule developmental biology. Plant Biotechnology Journal. 2017; 16:1363-1374.

23. Kim S. Park J, Yeom SI, Kim YM, Seo E, Kim KT, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biology.2017; 18(1): 210.
24. Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. Nature Plants. 2018; 4:473-84.
25. Liu Z, Song T, Zhu Q, Wang W, Zhou J, Liao H. *De novo* assembly and analysis of *Cassia tora* seed transcriptome to identify genes involved in the biosynthesis of active metabolites. Bio Biotechnol Biochem. 2014; 78(5):791-99.
26. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012; 58:268-276.
27. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology. 2015; 16:1-11.
28. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JQ, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol. 2013; 31:1119-25.
29. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and conformational capture enable *de novo* mammalian reference genomes. Nature Genetics. 2016; 49:643-50.
30. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 2016; 13:1050.
31. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017; 356:92-95.
32. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nature Genetics. 2017; 49:643-650.
33. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. BMC Biol. 2017; 15:74.
34. Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. Genome size diversity and its impact on the evolution of land plants. Genes. 2018; 9:88.
35. Macas J, Novák P, Pellicer J, Čížková J, Koblížková A, Neumann P, et al. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PLoS ONE. 2015; 10:e0143424.
36. English AC, Richards S, Han Y, Wang M, Vee V, Qu J,... Gibbs RA.. Mind the gap: Upgrading genomes withPacific Biosciences RS long-read sequencing technology. PloS One, 2012,7, e47768.
37. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, & Shendure J.. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nature Biotechnology, 2013, 31: 1119–1125.

38. Li H. & Richard D.. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009, 25, 1754-1760
39. Simao FA, Waterhouse RM, Panagiotis I, Evgenia VK, & Evgeny MZ. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015, **31**: 3210–3212..
40. Xu Z. & Wang H. LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research. 2007, **35**: 265-268.
41. Han YJ. & Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research.2010,* **38**:199.
42. Price A L, Jones NC. & Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics.2005*,* **21**: 351-358.
43. Edgar RC. & Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005, **21**, 152-158.
44. Hoede C, Arnoux S, Moissette M, Chaumier T, Inizan O, Jamilloux V, & Quesneville H. PASTEC: An automatic transposable element classification tool. PLoS One, 2014, 9(5), e91929.
45. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, & Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic acids research. 2016,**44**, 89-89.
46. Burge C. & Karlin S. Prediction of complete gene structures in human genomic DNA. Journal of molecular biology. 1997, **268**, 78-94.
47. Stanke M. & Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003, **19**, 215-225.
48. Majoros WH, Pertea M. & Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 2004,**20**, 2878-2879.
49. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Current protocols in bioinformatics 2007, 17, 4.3.1-4.3.28
50. Korf I. Gene finding in novel genomes. *BMC bioinformatics*. 2004, 5:59.
51. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008; 9: R7.
52. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Gonzales NR. CDD: A Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research, 2011, 39, D225–D229.
53. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, ... Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic acids research.2001, **29**, 22-28.
54. Kanehisa M, & Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 2000, 28, 27–30.

55. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin M J, ... Apweiler R. The UniProtGO annotation database in 2011. Nucleic Acids Research, 2012, 40, 565–570.
56. Zdobnov EM, & Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 2001,17, 847–848.
57. Bairoch A. PROSITE: A dictionary of sites and patterns in proteins. Nucleic Acids Research,1991, 19, 2241–2245.
58. Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, ... Bairoch A. HAMAP: A database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Research, 2009, 37, 471–478.
59. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, ... Bateman A. Pfam: Clans, web tools and services. Nucleic Acids Research, 2006, 34, 247–251.
60. Attwood TK, Beck ME, Bleasby A J, & Parry-Smith DJ. Prints-a protein motif fingerprint database. Protein Engineering, 1994, 7, 841–848
61. Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D: The ProDom database of protein domain families: more emphasis on 3D. *Nucleic acids research* 2005, **33**:D212-D215.
62. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, ... Bork P. SMART 4.0: Towards genomic data integration. Nucleic Acids Research, 2004, 32, 142–144
63. Haft DH, Selengut JD, & White O. The TIGRFAMs database of protein families. Nucleic Acids Research, 2003, 31, 371–373.
64. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, ... Kourtesis P. PIRSF: Family classification system at the Protein Information Resource. Nucleic Acids Research, 2004, 32, 112–114.
65. Ghosh S, & Chan CK. Analysis of RNA-seq data using Tophat and Cufflinks. Methods in Molecular Biology, 2016, 1374, 339–361.
66. Lees J, Yeats C, Perkins J, Sillitoe I, Rentzsch R, Dessailly BH, & Orengo C. Gene3D: A domain-based resource for comparative genomics, functional annotation and protein network analysis. Nucleic Acids Research, 2012, 40, 465–471.
67. Li L, Stoeckert CJ, & Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Research, 2003, 13, 2178–2189.
68. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, & Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology, 2010, 59, 307–321.
69. De Bie T, Cristianini N, Demuth JP, & Hahn MW. CAFE: A computational tool for the study of gene family evolution. Bioinformatics, 2006, 22, 1269–1271.

Figures

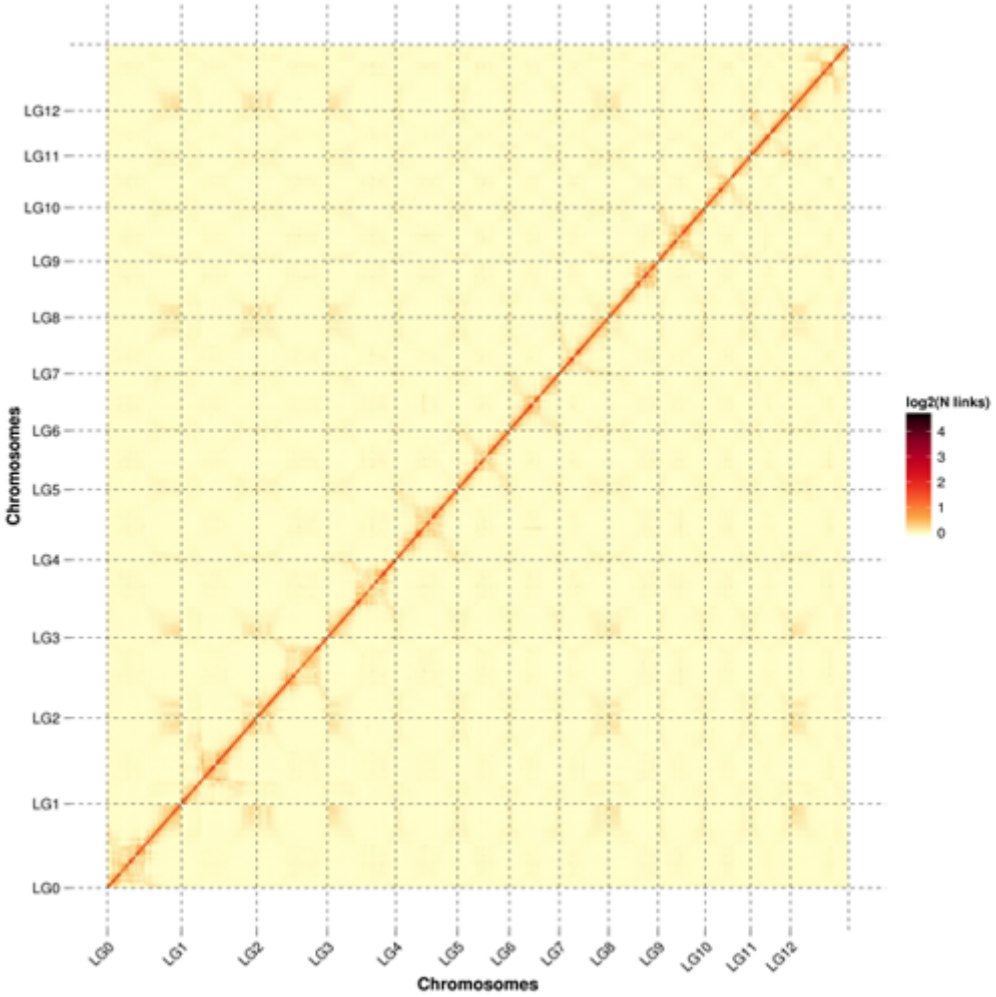

Figure 1

Contact matrices generated by aligning the Hi-C dataset to the final genome assembly. LG0–LG12: Lachesis Groups 0–12. Abscissa and ordinate represent the order of each bin in its corresponding chromosome group.

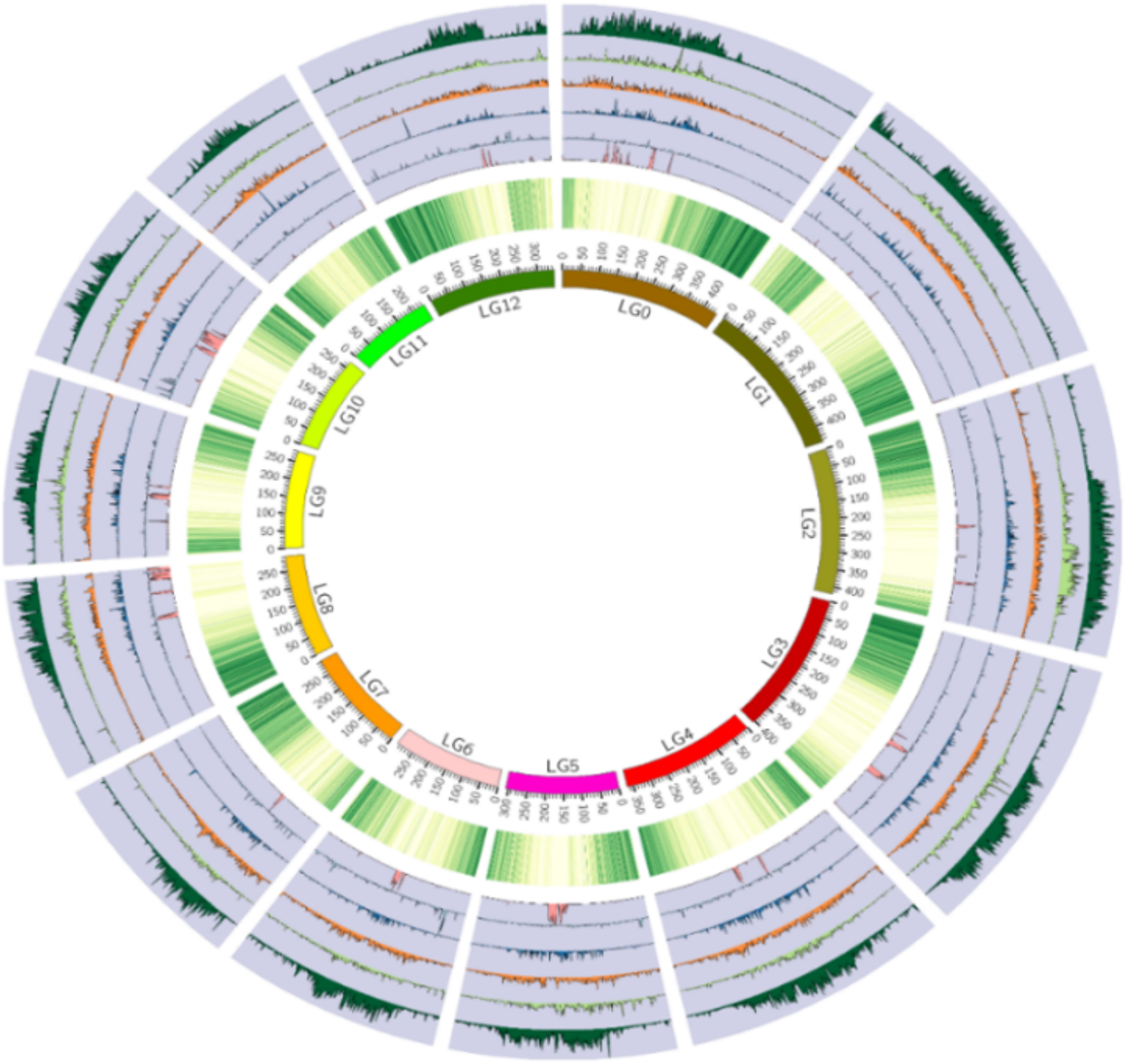

Figure 2

Cassia tora genome comprising 13 chromosomes (LG0–12) scaled according to assembly length. Tracks from outside to inside show Gypsy-type LTRs (48,574), LARD-type LTRs (16,354), Copia-type LTRs (18,169), CACTA repeats (20,203), Helitron-type elements (33,411), SSR markers (7,091), genes (32,361), and Lachesis groups 0–12. Window size = 100 kb.

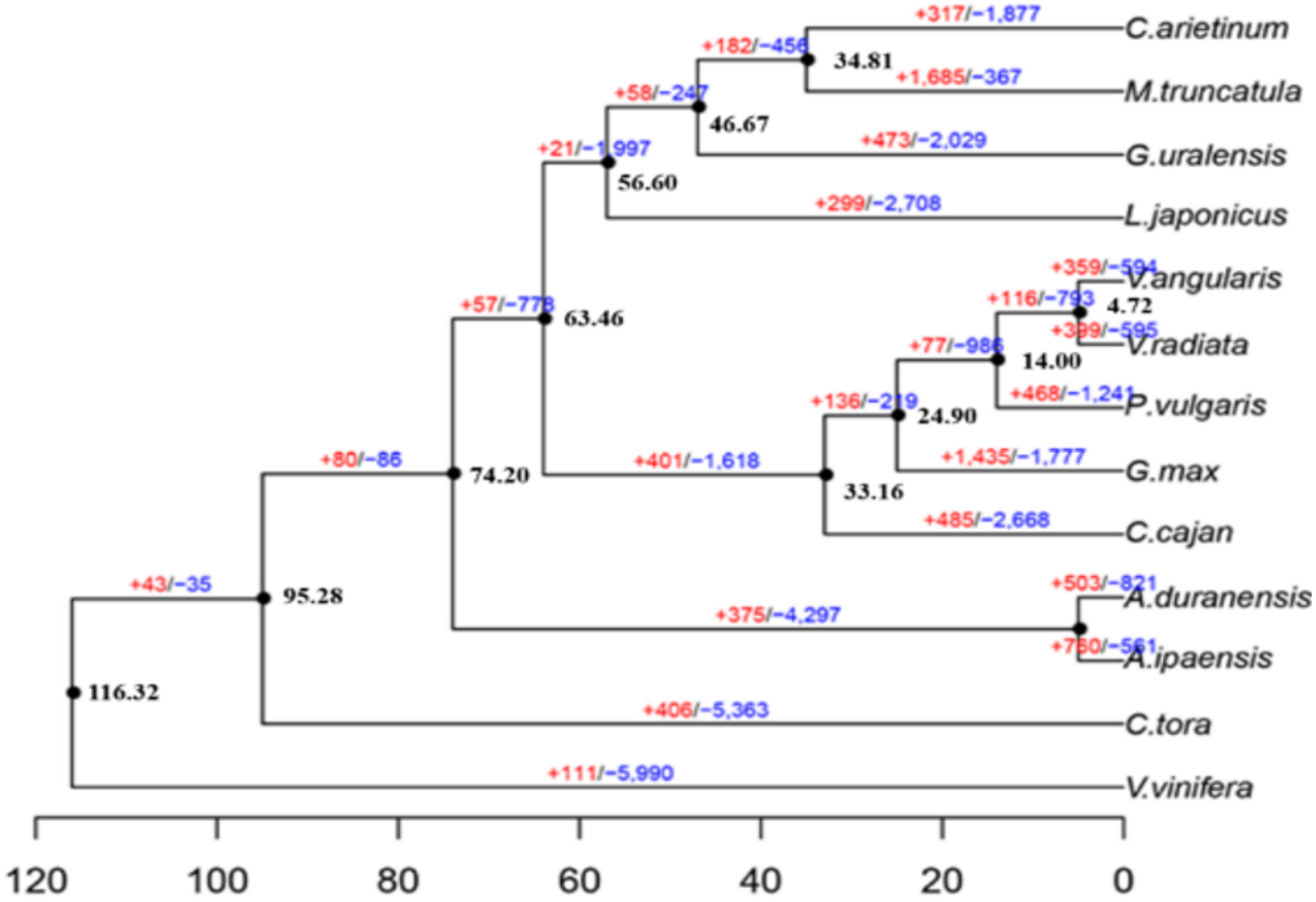

Figure 3

Neighbor-joining (NJ) tree for C. tora and 11 other legume species. Phylogenetic analysis and divergence time estimation for 13 plant species and gene family dynamics for each branch. Divergence times are indicated by black numbers next to branching nodes (right). Blue and red numbers below each species name (left) indicate numbers of gene family contraction and expansion events, respectively.

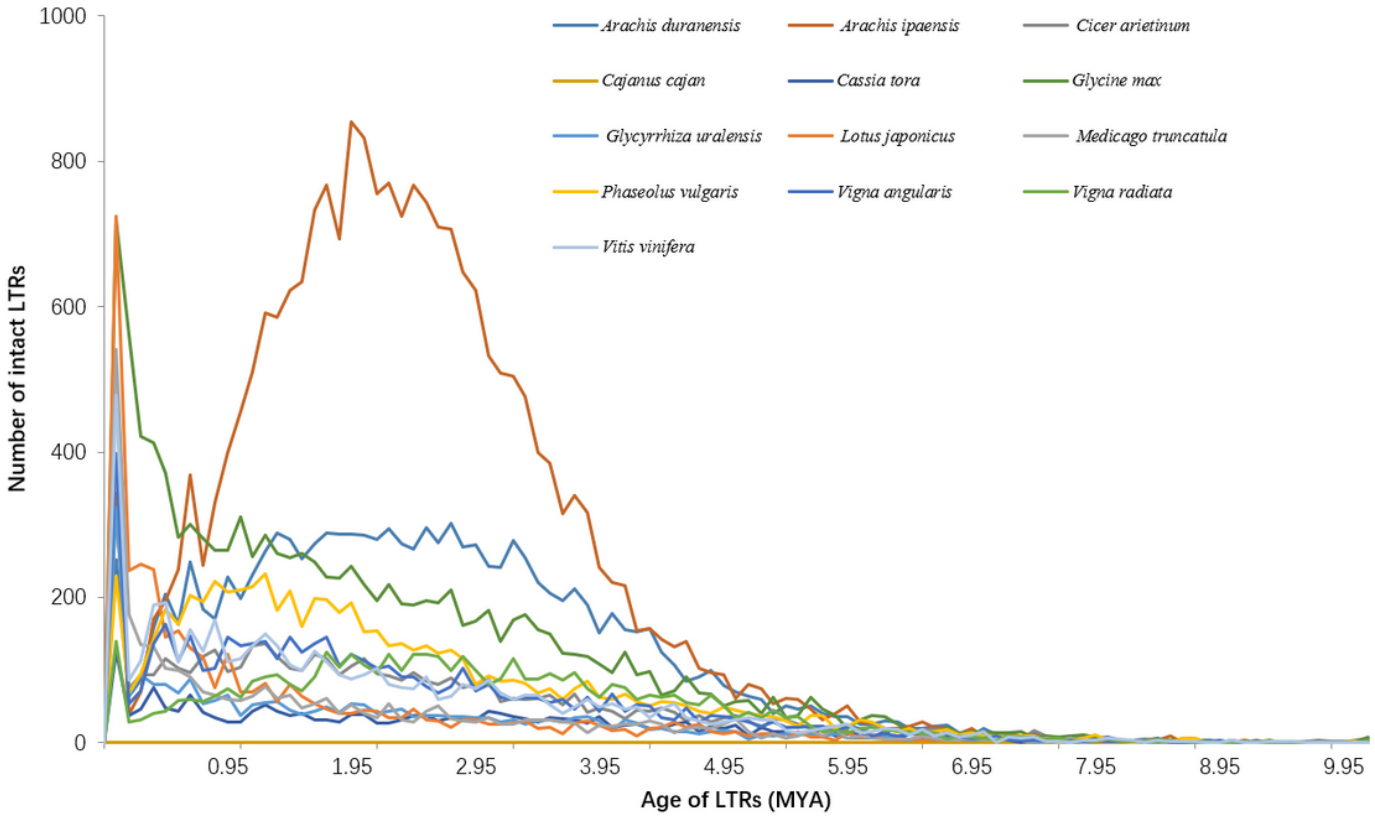

Figure 4

Analysis of LTR retrotransposon number and insertion time in C. tora. Curves represent accumulated total number of LTRs throughout evolution. Cto, Cassia tora; Van, Vigna angularis; Vra, Vigna radiata; Pvu, Phaseolus vulgaris; Gur, Glycyrrhiza uralensis; Gly, Glycine max; Lco, Lotus japonicus; Mtr, Medicago truncatula; Car, Cicer arietinum; Cca, Cajanus cajan; Adu, Arachis duranensis and Aip, Arachis ipaensis (both diploid ancestors of cultivated peanut); Vvi, Vitis vinifera (non-legume species).

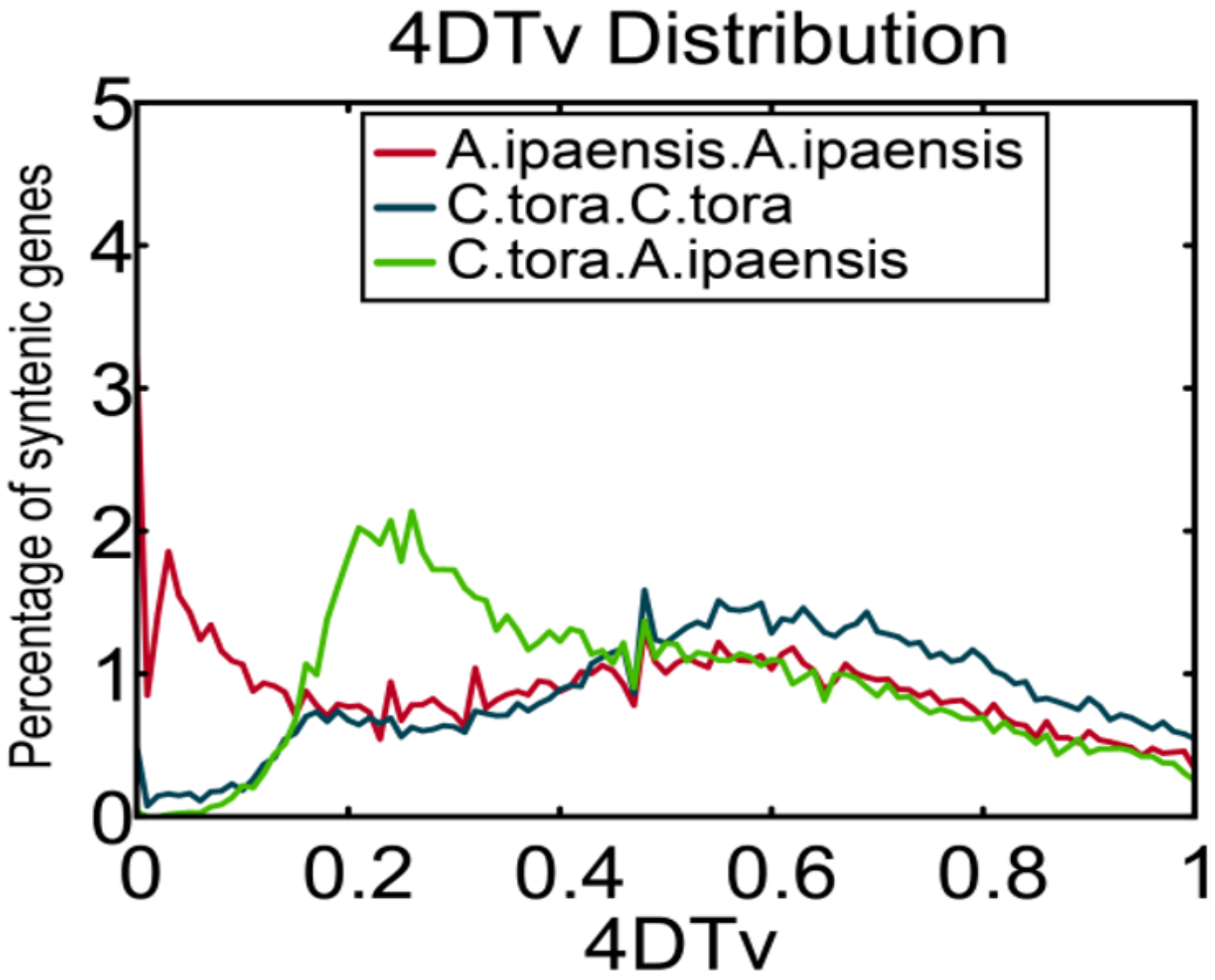

Figure 5

4DTv distribution between C. tora and A. ipaensis. X-axis shows mutation rates of genes homologous to 4DTv. Y-axis shows proportions of homologous genes exhibiting 4DTv.

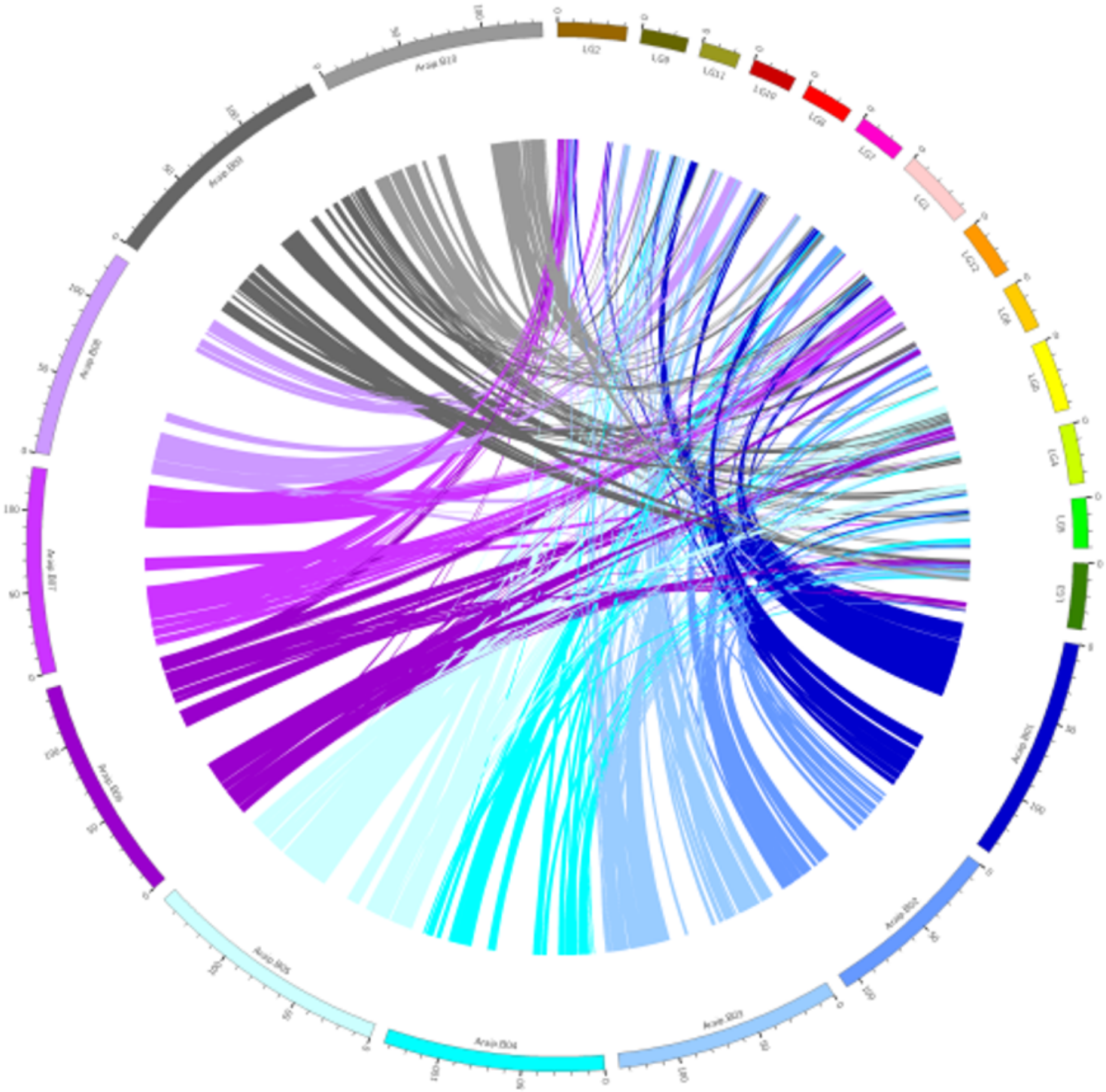

Figure 6

Collinearity analysis between C. tora and A. ipaensis. Circular representation of order of C. tora genome based on conserved synteny with peanut A genome. Corresponding homologs between C. tora and A. ipaensis are based on reciprocal best hit of proteins and are represented by colored lines corresponding to each peanut chromosome. Gene density is expressed per 100 kb.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalInformation12.21.docx