


Brief Communication

The chromosome-level genome sequence of the camphor tree provides insights into Lauraceae evolution and terpene biosynthesis

Tengfei Shen¹, Haoran Qi¹, Xiaoyue Luan¹, Wenlin Xu¹, Faxin Yu², Yongda Zhong^{2,*} and Meng Xu^{1,*} ¹Co-Innovation Center for Sustainable Forestry in Southern China, Key Laboratory of Forest Genetics and Biotechnology Ministry of Education, Nanjing Forestry University, Nanjing, China²The Key Laboratory of Horticultural Plant Genetic and Improvement of Jiangxi Province, Institute of Biological Resources, Jiangxi Academy of Sciences, Nanchang, China

Received 29 September 2021;

revised 28 October 2021;

accepted 8 November 2021.

*Correspondence (Tel +86-025-85427412; fax +86-025-85427412; emails xum@njfu.edu.cn or zhongyongda@jxas.ac.cn)

Keywords: genome assembly, evolutionary status, whole-genome duplication, terpenoid metabolic pathway, *Cinnamomum camphora*.

Cinnamomum camphora, commonly known as the camphor tree, an economically and ecologically important aromatic tree species, has a long history of cultivation and utilization. It is the representative species of subtropical evergreen broadleaved forests in eastern Asia and an important raw material for essential oil production worldwide. The whole camphor tree is rich in terpenoids, which are widely used in industrial and pharmaceutical applications. According to the main volatile components of leaf essential oils (LEOs), such as monoterpenes, sesquiterpenes and diterpenes, camphor trees can be subdivided into five chemotypes: the borneol, camphor, cineole, linalool and nerolidol types. However, the genetic bases for the biosynthesis of these components in camphor trees are not yet well understood. The camphor tree is a member of the Lauraceae family of Laurales, which comprises the magnoliids and three other related groups (Magnoliales, Canellales and Piperales). Despite several available magnoliid genomes (Chaw *et al.*, 2019; Chen *et al.*, 2019, 2020; Lv *et al.*, 2020), the evolutionary relationships among magnoliids, eudicots and monocots remain controversial (Qin *et al.*, 2021). We herein report the assembly of a high-quality reference genome for the camphor tree, which helps address the above-mentioned problems.

The camphor tree is diploid ($2n = 24$) with an estimated haploid genome size of approximately 785 Mb, as determined using 17-mer analysis of $180\times$ Illumina reads. The genome was initially assembled by hifiasm v0.13 with 1 267 672 PacBio high-fidelity long reads (HiFi reads, N50 = 16.1 kb), and further scaffolding was combined with 122.72 Gb of reads from chromosome conformation capture. Contigs were anchored and oriented on 12 pseudochromosomes using 3d-dna, generating chromosome-level sequences of 670.29 Mb, with a contig N50 value of 2.41 Mb and a scaffold N50 of 60.19 Mb (Figure 1a,b). BUSCO analysis showed that the completeness of the camphor tree genome was 95.2%, and the LTR assembly index (LAI) also had a high score (18.2), indicating the excellent

continuity of the assembly. The camphor tree genome harbours 361.82 Mb of repetitive sequences, of which long terminal repeat (LTR) retrotransposons accounted for 27.66% of the whole genome. The *gypsy* and *copia* elements were the predominant LTRs, occupying 22.56% of the *C. camphora* genome, which is between the values for *Litsea cubeba* (45.31% in a 1325.69 Mb genome) and *Cinnamomum kanehirae* (16.50% in a 730.7 Mb genome).

Through *ab initio* modelling, protein-based searches and transcript analysis of long-read isoform sequencing and short-read RNA sequencing data, a high-confidence set of 29 919 protein-coding gene models (concealing 37 295 protein-coding transcripts) was predicted by Maker2 in the *C. camphora* genome and was located on the 12 pseudochromosomes. Of these protein homologs in the TrEMBL database, 92.48% and 59.71% could be assigned Gene Ontology terms. The proteome of these protein-coding genes was estimated to be 90.8% complete based on BUSCO analysis, which is slightly higher than the values for the other two related species, namely, *L. cubeba* (89.2%) and *C. kanehirae* (89%). Six magnoliids share 8276 gene families containing 18 044 genes, of which 127 families (174 genes) were unique to *C. camphora* and were significantly enriched in ascorbate and aldarate metabolism ($9.39E-05$), monoterpene biosynthesis ($1.40E-04$), glutathione metabolism ($1.48E-04$) and so on.

The *C. camphora* genome, with superior contiguity and reliable annotations compared with the other published Lauraceae genomes, shed light on the mysteries of magnoliid evolution. The 104 strictly single-copy ortholog sets derived from seven magnoliids, five eudicots, six monocots and two out-group species were used to reconstruct high-confidence phylogenetic trees by protein and nucleotide sequence alignments. Similar topologies strongly support Lauraceae, representing magnoliids as the sister lineage to eudicots. Using MCMCTree with fossil calibrations, the separation between *C. camphora* and *C. kanehirae* was found to have occurred approximately 4.75 Mya, and the divergence time of magnoliids and eudicots was ~144.26 Mya (Figure 1c). During evolution, 1110 gene families in the camphor tree underwent expansion, while 1528 gene families underwent contraction. Intriguingly, 2169 expanded genes belonging to 163 rapidly evolving families were significantly enriched in monoterpene biosynthesis ($3.90E-17$), spliceosome ($4.47E-15$) and ABC transporters ($2.16E-12$), which suggests that chemotype diversification in the camphor trees may be promoted by species-specific genes and tachytelic gene

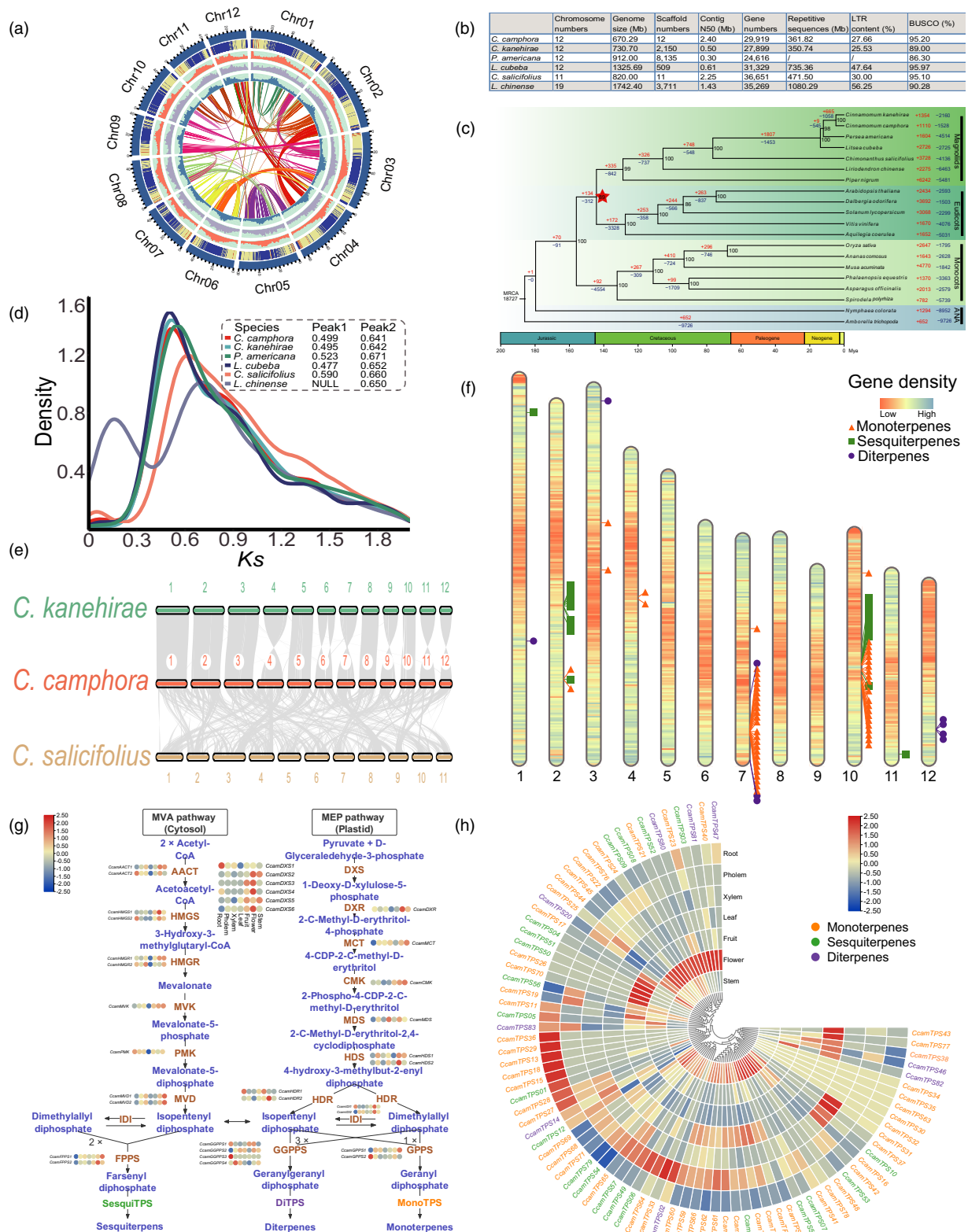


Figure 1 (a) Genome features across 12 chromosomes. The circular map shows, from outside to inside, ideograms of the 12 chromosomes, density of genes (blue-red scale), density of LTRs, density of *Copia*, density of *Gypsy* and syntenic blocks. (b) Statistics for the assembly and annotation of the six published magnoliid genomes. (c) Dated phylogeny for 20 plant species with ANA as an out-group, a time scale is shown at the bottom. The bootstrap value is given in black. The gene families that expanded and contracted are given in red and blue, respectively. (d) Density distribution of *Ks* for paralogous gene pairs of the six magnoliid genomes. (e) Interspecific collinearity at the chromosome level among *C. camphora*, *C. kanehirae* and *C. salicifolius*. The grey line connects matched gene pairs. (f) Distribution pattern of 83 *TPS* genes on chromosomes. (g) Key genes involved in terpenoid backbone biosynthesis pathways in the camphor tree genome. (h) Transcriptional heatmap of 83 *CcamTPS* genes.

families related to terpenoid biosynthesis. Comparative genome analysis revealed that there were 68 positively selected genes in the camphor tree.

Genome collinearity and distinctions of synonymous substitutions per synonymous site (K_s) revealed that the evolutionary trajectory of the camphor tree genome has been generally shaped by whole-genome duplication (WGD) events. Intragenomic synteny examination revealed a total of 385 syntenic blocks that contained 20 684 collinear pairs in the camphor tree genome assembly, and the degree of interspecific collinearity between the camphor tree and its related species was consistent with their evolutionary topologies (Figure 1a,e). By estimating intragenomic and interspecies K_s distributions, two signature peaks for WGD events were observed at $K_s \approx 0.499$ and 0.641 for six species (Figure 1d), showing that after the distant WGD event (ϵ) encountered by all extant angiosperms, a recent WGD (~ 76 Mya) in *C. camphora* was shared by all the Lauraceae species and an ancient WGD event (~ 124 Mya) arose before the divergence of Magnoliales and Laurales.

Terpene synthases (TPSs) are critical rate-limiting enzymes that produce bioactive terpenoids with multifarious backbones. Compared with the 76 TPS genes in the 12 pseudochromosomes of the stout camphor genome, a total of 83 *CcamTPS* genes were predicted and annotated in the camphor tree genome, including 53 monoTPSs, 21 sesquiTPSs and nine diTPSs. These rapidly evolving genes were distributed unevenly on seven chromosomes and were clustered together in tandem (Figure 1f). MonoTPS genes were concentrated in the middle region of Chr7 and Chr10, and sesquiTPSs were mainly distributed in the middle region of Chr2 and Chr10. Tandem rearrangement of TPS genes may be associated with the mass production of terpenoids in the genus *Cinnamomum*. The plant TPS family is divided into seven subfamilies, of which the TPS-d subfamily is specific to gymnosperms. The camphor tree genome has six subfamilies, including 20 TPS-a, 42 TPS-b, 1 TPS-c, 10 TPS-e/f and 10 TPS-g members, and the TPS-a and TPS-b subfamilies are the most diverse, presumably contributing to the biosynthesis of monoterpenes and sesquiterpenes. Transcriptome sequencing showed the tissue-specific expression profiles of 83 TPS genes (Figure 1g,h), of which five monoTPSs were not expressed in seven tissues. From these results, we inferred that the rapid expansion, tandem arrangement and tissue-specific expression of terpene biosynthetic genes powered the chemotypic diversification of the camphor tree.

In summary, the reference-quality genome of *C. camphora* provides new insights into terpene biosynthesis and lays the foundation for better elucidating the evolution and diversification of Lauraceae.

Acknowledgements

This work is supported by grants from the National Natural Science Foundation of China (Grant Nos. 32060354, 31860079 and 32160397) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). We are grateful to Professor Meng-li Xi for performing the karyotype analysis and Mr. Yin-Cong Gu and Ms. Dong An (Shanghai OE Biotechnology Co., Ltd.) for their technical support in genome data analysis.

Conflicts of interest

The authors declare no conflicts of interest.

Author contributions

M.X., Y.Z and F. Y. conceived the project. M.X., T.S., Y.Z., H.Q., X.L. and W.X. participated in the data analysis. M.X. and T.S. wrote the manuscript.

References

- Chaw, S.-M., Liu, Y.-C., Wu, Y.-W., Wang, H.-Y., Lin, C.-Y., Wu, C.-S., Ke, H.-M. et al. (2019) Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nature Plants*, **5**, 63–73.
- Chen, J.H., Hao, Z.D., Guang, X.M., Zhao, C.X., Wang, P.K., Xue, L.J., Zhu, Q.H. et al. (2019) Liriodendron genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nature Plants*, **5**, 328.
- Chen, Y.-C., Li, Z., Zhao, Y.-X., Gao, M., Wang, J.-Y., Liu, K.-W., Wang, X. et al. (2020) The *Litsea* genome and the evolution of the laurel family. *Nat. Commun.* **11**, 1675.
- Lv, Q.D., Qiu, J., Liu, J., Li, Z., Zhang, W.T., Wang, Q., Fang, J. et al. (2020) The *Chimonanthus salicifolius* genome provides insight into magnoliids evolution and flavonoids biosynthesis. *Plant J.* **103**, 1910–1923.
- Qin, L., Hu, Y., Wang, J., Wang, X., Zhao, R., Shan, H., Li, K. et al. (2021) Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome. *Nature Plants*, **7**, 1239–1253.