scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the traditional medicinal plant *Lindera aggregata*

Yujie Shi[1], Zhen Chen[1], Junxia Ge[2], Jingyong Jiang[3], Qianfan Li[4], Yiluo Lin[2], Weifu Yu[2] & Wei Zeng[1] ✉

Lindera aggregata **is a renowned medicinal plant in China, particularly the variety from Tiantai, Zhejiang Province, which is esteemed for its superior medicinal properties. Beyond its medicinal value, it holds significant economic potential and phylogenetic significance. Utilizing a range of sequencing techniques, we have successfully assembled and annotated a high-quality chromosome-level genome of *L. aggregata*. The assembled genome spans approximately 1.59 Gb, with a scaffold N50 length of 132.62 Mb. Approximately 93.07% of the assembled sequences have been anchored to 12 pseudo-chromosomes, and 70.02% of the genome consists of repetitive sequences. According to the annotations, a total of 33,283 genes are identified, of which 96.95% can predict function. This high-quality chromosome-level assembly and annotation will greatly assist in the development and utilization of *L. aggregata*'s valuable resources, and also provide a crucial molecular foundation for investigating the evolutionary relationships within the Lauraceae family and the mechanisms behind the synthesis of active ingredients in *L. aggregata*.**

Background & Summary

The Lauraceae family encompasses approximately 100 species of the genus *Lindera*, which are extensively distributed across tropical, subtropical, and temperate zones in Asia and the central region of the United States[1]. Over 40 species are found in China, representing roughly 46% of the genus's total[2]. Many species within this genus possess aromatic oils that are utilized for culinary spices and medicinal applications[3]. Their seeds, abundant in fats, can be processed into soap and lubricants[4]. Additionally, certain tree species of this genus yield wood that is suitable for construction materials or furniture-making[1].

Lindera aggregata (Sims) Kosterm (called "Wu-Yao" in Chinese), a traditional medicinal plant in China, is predominantly cultivated in regions such as Zhejiang, Jiangxi, Fujian, Anhui, Hunan, Guangdong, and Guangxi[3]. Notably, the *L. aggregata* originating from Zhejiang, particularly the renowned "Tiantai Wu-Yao", is esteemed for its superior quality[5]. As one of the famous "New Zhejiang eight traditional Chinese medicine", it holds a prestigious status all over the world. Historical texts from ancient Chinese classics, including "Ben Cao Meng Quan" and "Ben Cao Gang Mu" indicate that the jointed tuberous roots of *L. aggregata* were the primary medicinal components (Fig. 1A), rather than the taproots[6]. In Traditional Chinese Medicine, qi is considered an extremely subtle yet potent force that circulates continuously within the human body[7]. It plays a crucial role in promoting and regulating metabolism and maintaining the body's life processes[8]. *L. aggregata* is known for its properties to activate qi, alleviate pain, warm the kidneys, and dispel cold[9]. It is utilized to address symptoms such as abdominal pain, dysmenorrhea, frequent urination, rheumatism, and indigestion[10]. Due to its substantial medicinal values and extensive pharmacological effects, this plant has garnered increasing attention in recent years.

Over the past few decades, researchers had delved into the *L. aggregata* from various angles, encompassing chemical analysis, pharmacological mechanisms, and quality control methodologies. So far, more than 260 compounds have been isolated and identified from this species, including flavonoids, alkaloids, terpenes, volatile oils, etc[2]. Notably, the "China Pharmacopeia" has utilized linderane and norisoboldine as chemical

[1]Zhejiang Provincial Key Laboratory of Plant Evolutionary Ecology and Conservation, College of Life Sciences, Taizhou University, Taizhou, 318000, China. [2]Zhejiang Hongshiliang Group Tiantai Mountain Wu-Yao Co., Ltd., Taizhou, 318000, China. [3]Institute of Horticulture, Taizhou Academy of Agricultural Sciences, Linhai, 317000, China. [4]State Key Laboratory of Subtropical Silviculture, College of Forestry and Biotechnology, Zhejiang A&F University, Hangzhou, 311300, China. ✉e-mail: zengw@tzc.edu.cn

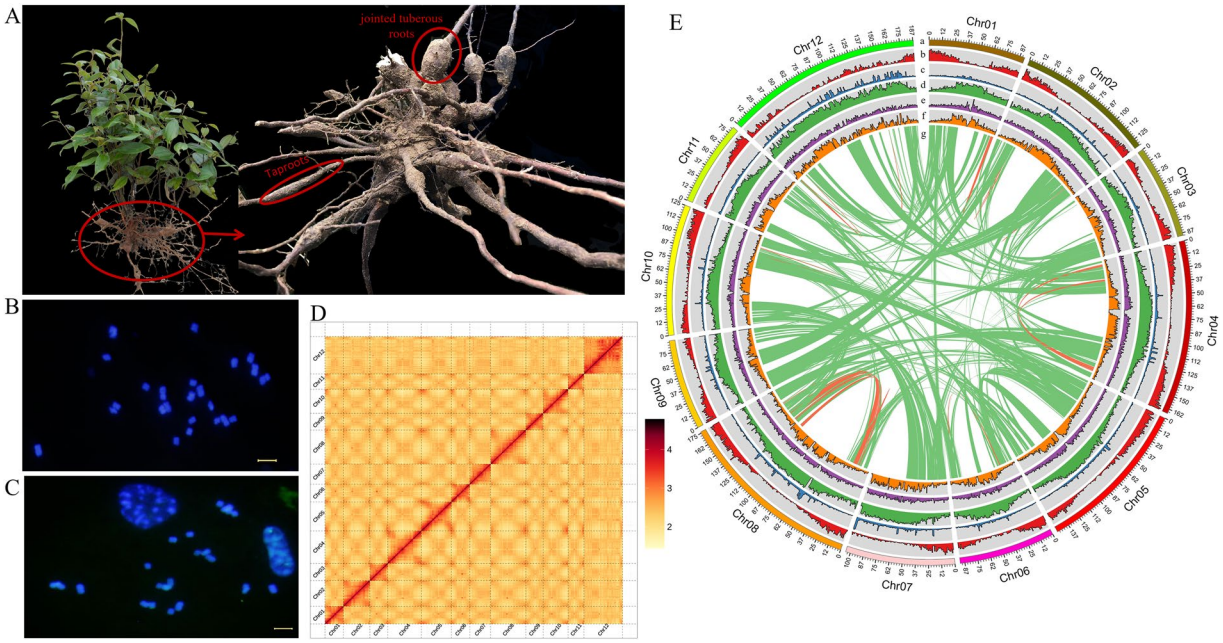

Fig. 1 The morphological and genome characteristics of *Lindera aggregata*. (**A**) The whole plant and roots of *L. aggregata*, in which the jointed tuberous roots and the taproots are circled in red. (**B**) Chromosomes fluorescence staining of *L. aggregata*, with a scale of 5 um. (**C**) Fluorescence *in situ* hybridization of chromosomes, with a scale of 5 um. (**D**) Hi-C interaction heat map. (**E**) Circle plot of genome assembly and annotation. a, Chromosome-scale pseudomolecules (chr01-chr12); b-f, Distribution of gene density, GC density, transposon element, copia and gypsy LTR density, respectively; g, Colinearity blocks in genome.

Parameter	Genome
Genome size	1,595,824,669 bp
GC content	40.39%
Contig number	1,877
Contig N50	12,106,626 bp
Contig N90	1,826,541 bp
Scaffold number	1,659
Scaffold N50	132,620,616 bp
Scaffold N90	81,699,946 bp
Chromosome number	12
Chromosome length	1,485,206,438 bp (93.07%)
Mitochondria length	912,473 bp (0.06%)
Chloroplast length	154,736 bp (0.01%)

Table 1. Summary of *Lindera aggregate* genome assembly.

markers for evaluating the quality of *L. aggregata* roots[6]. However, due to the absence of a complete genome for *L. aggregata*, it is currently not feasible to analyze the synthetic pathways of the primary chemical constituents and the mechanisms underlying pharmacological effects at the molecular level. This significantly hampers the advancement and practical application of *L. aggregata*. Consequently, the acquisition of high-quality reference genomes is essential for enhancing the utilization of resources in *L. aggregata* and for investigating the phylogeny of Lauraceae plants.

In this study, we employed PacBio HiFi reads (91.80 Gb, 55×), Illumina reads (101.13 Gb, 60×), Hi-C reads (236.46 Gb), and RNA-seq data (39.40 Gb) to assemble and annotate the genome of *L. aggregata*. The assembled genome of *L. aggregata* spans 1.59 Gb, anchored on 12 pseudo-chromosomes, with a contig N50 of 12.11 Mb, and an assembly completeness of 93.07% (Table 1; Fig. 1D,E; Table S1). Based on BUSCO assessment, the proportion of complete core genes was 97.3%, and the proportion of missing genes was 1.8%, indicating that the assembly completeness was good (Fig. 2; Table S7). The sizes of chloroplast and mitochondrial genomes were 154,736 bp and 912,473 bp, respectively. A total of 1,832,346 repeat sequences were identified, with a cumulative length of approximately 1.12 Gb, which accounts for 70.02% of the assembled genome. Long terminal repeats (LTRs) constitute the largest proportion, with 667,992 sequences and a cumulative length of 602.85 Mb, representing 37.78% of the entire genome. Compared with other Lauraceae species, the

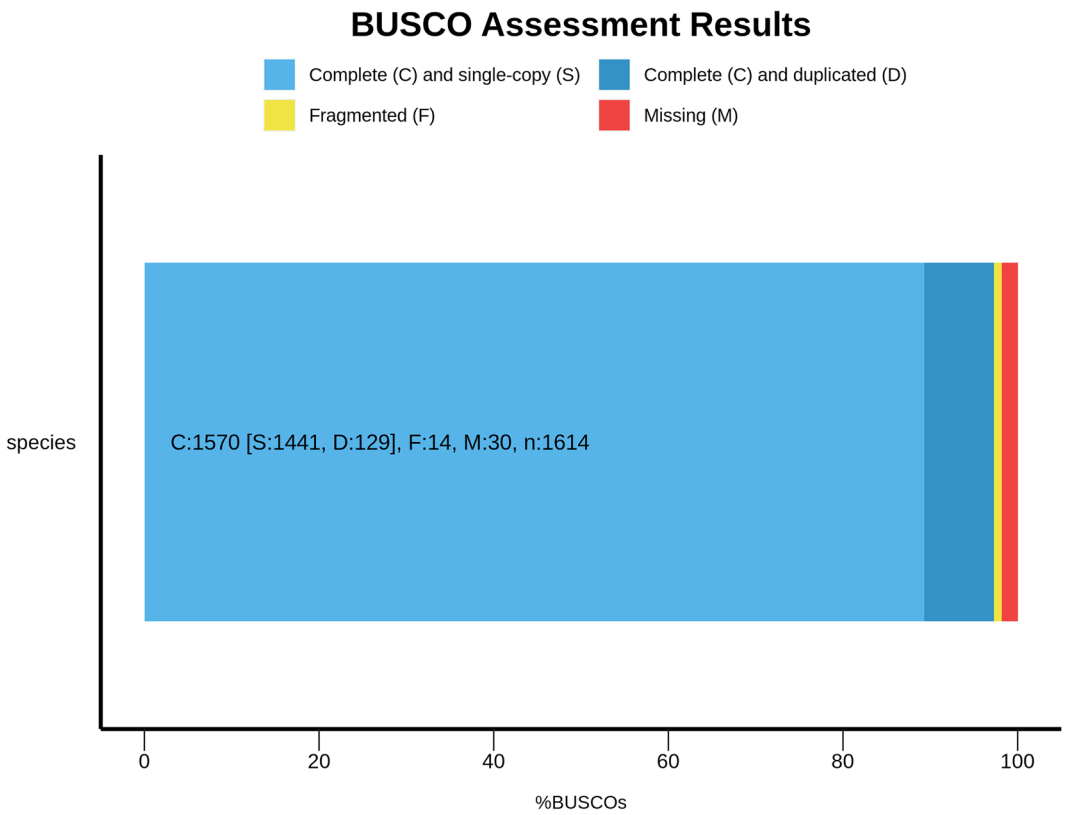

Fig. 2 Evaluation of genomic integrity of *Lindera aggregate* based on BUSCO.

Feature	Number
Repeat sequence	1,832,346
Gene	33,283
Protein-coding gene	32,268
ncRNA	38,108

Table 2. Summary of *Lindera aggregate* genome annotation.

size of *L. aggregate* genome was larger, only smaller than that of *Lindera glauca* (2.09 Gb), but the repetitive sequences proportion of *L. aggregate* genome was the largest, indicating that its genome was more complex (Table S15). Moreover, we predicted 33,283 genes within the genome, of which 96.95% were assigned putative functions (Table 2), similar in number to other Lauraceae plants (ranging from 28,198 to 48,139, Table S15). The successful construction of high-quality reference genomes will enhance our understanding of the evolutionary relationships among Lauraceae plants and further facilitate research into the regulation of medicinal active ingredients in *L. aggregata*, as well as molecular-assisted breeding efforts.

Methods

Samples collection. In order to extract genomic DNA, fresh young leaves of *L. aggregate* were collected from an adult plant at the "Wuyao" planting base in Sanzhou Township, Tiantai County, Zhejiang Province, China (120°48.23′E, 29°12.71′N). Furthermore, young leaves, fruits, stems, flower buds, taproots and jointed tuberous roots were collected from the same plant for subsequent transcriptome sequencing. After all samples were collected, they were quickly frozen in liquid nitrogen and then stored in a −80 °C refrigerator. DNA and RNA extraction and sequencing were completed by Novogene Biotechnology Co., Ltd. in Tianjin, China.

Library construction and genome sequencing. Total DNA was extracted from young leaves of *L. aggregate* by a modified CTAB method[11]. The concentration of DNA was assessed using Nanodrop and Qubit fluorometer. The integrity and purity of DNA were assessed using 1% agarose gel electrophoresis. Short-read WGS sequencing data was obtained through the Illumina NovaSeq X plus platform, resulting in approximately 101.13 Gb of raw data (Table S2).

Prior to HiFi sequencing, DNA samples were first subjected to agarose gel electrophoresis to assess their quality (main band >30 kb). High-quality DNA samples that pass the test were selected and randomly interrupted by a Covaris ultrasonic disruptor to bring the fragment size within the target range of 15–18 kb, and

large fragments of DNA were purified using magnetic beads for enrichment. Subsequently, each PacBio single molecule real-time (SMRT) library was built using SMRTbell Express Template Prep Kit 2.0. The library was then sequenced in CCS mode on the Pacific Bioscience Revio platform, and high-fidelity (HiFi) reads were generated using Circular Consensus Sequencing (CCS) workflow v8.0.1. This process produced a total of 91.80 Gb data, with an average read length of approximately 17.22 kb, and an N50 read length of approximately 17.23 kb (Table S3).

To prepare a Hi-C library, cells were treated with formaldehyde to bind DNA and protein to be fixed. After cell lysis, the cross-linked DNA was treated with restriction enzymes to create gaps on both sides of the cross-linking point. During end repair, biotin-labeled ends of oligonucleotides were added. Subsequently, adjacent DNA fragments were ligated by treatment with nucleic acid ligase. The proteins at the junction were digested with protease to uncross-link the protein and DNA. Then genomic DNA was extracted and the DNA was randomly broken into fragments with a length of 350 bp using a Covaris crusher for recovery. Finally, sequencing was performed on the Illumina NovaSeq X plus platform in the PE150 model (Table S4).

Transcriptome sequencing. Total RNAs were isolated from the young leaves, fruits, stems, flower buds, taproots and jointed tuberous roots of the same plant. The concentration of RNAs were assessed by Nanodrop and Agilent 2100. The integrity and purity of RNAs were assessed by agarose gel electrophoresis and Agilent 2100. After the RNA samples were tested, mRNA was enriched with magnetic beads with Oligo (dT). Subsequently, fragmentation buffer was added to break the mRNA into short fragments. Using the mRNA as a template, one-strand cDNA was synthesized using six-base random hexamers. Then buffer, dNTPs, and DNA polymerase I and RNase H were added to synthesize two-strand cDNA. The double-strand cDNA was then purified using AMPure XP beads. The purified double-stranded cDNA was first end-repaired, A-poly, and connected to sequencing adapters, and then fragment size was selected using AMPure XP beads. PCR amplification was then carried out and the PCR products were purified using AMPure XP beads to obtain the final sequencing library. Sequencing was performed via the Illumina NovaSeq X plus platform. Finally, a total of 40.20 Gb RNA-Seq data (Table S5) was obtained for subsequent annotation.

Genome size estimation. The frequency of 17-kmers was generated based on clean WGS reads via the jellyfish v2.2.7 tool[12] and the characteristics of genome were evaluated via GenomeScope[13]. The estimated genome size was approximately 1,679.58 Mb, the heterozygosity rate was approximately 0.97%, and the repeat sequence proportion was approximately 72.64% (Table S6). Moreover, hybrid k-mer pairs were extracted from k-mer data through Smudgeplot v0.4.0 software[14] and the hybrid k-mer pairs were trained. Then, the total and relative coverage of k-mer pairs were compared, and the number of heterozygous k-mer pairs was counted to infer that the genome of *L. aggregate* was diploid (Figure S2).

Chromosome karyotype analysis. The seedlings of *L. aggregate* were cultured to obtain roots with active meristems. Dinitrogen oxide was used to treat the cells and induce them to mitosis to obtain a large number of metaphase cells. After DAPI staining, clear and intuitive chromosomes were obtained through high-resolution fluorescence microscopy and CCD imaging equipment. Moreover, fluorescent probes based on telomeres and conserved repeats of 18SrDNA were used to conduct fluorescence *in situ* hybridization on the samples to determine the chromosome ploidy characteristics of the species. The results showed that the number of chromosomes was 24, and 2 chromosomes showed strong hybridization signals, suggesting that the sample was diploid (2n = 2x = 24; Fig. 1B,C).

Genome assembly. PacBio HiFi reads were assembled into contigs using Hifiasm v0.19.5[15] with parameter (−l = 2, -n = 4). The assembled draft genome of *L. aggregate* was 1,595.82 Mb, with contig N50 sizes of 12.11 Mb (Table S1). Subsequently, based on the Hi-C data obtained by sequencing, the assembled contigs/scaffolds sequences were mounted to the near-chromosome level using ALLHIC v0.9.8 software[16] (enz = DpnII, CLUSTER = n), including chromosome clustering, orientation, and sorting. Manual inspection and adjustment were performed using Juicebox v1.11.08 software[17] (pre -n -q 0 or 1), primarily focusing on refining chromosome segment boundaries and correcting assembly errors. Finally, the genome was obtained at the chromosome level, the size was 1,485.21 Mb, with 93.07% of sequences anchored to 12 pseudochromosomes and a scaffold N50 size of 132.62 Mb (Table 1; Fig. 1D,E; Table S1). Additionally, the Getorganelle v1.7.7[18] and PMAT v1.5.3[19] were used to assemble chloroplast and mitochondrial genomes with default parameters, respectively.

Genome annotation. Genome annotation mainly includes repeat sequence annotation, gene structure annotation, gene function annotation and non-coding RNA annotation. A combined strategy based on homology alignment and de novo search to identify the whole genome repeats were applied in our repeat annotation. The homolog prediction commonly used Repbase database[20,21] employing RepeatMasker v4.1.2-p1 software and its in-house scripts (RepeatProteinMask) with default parameters to extracted repeat regions[22]. And ab initio prediction built de novo repetitive elements database by RepeatModeler v2.0.3[23] with default parameters, then all repeat sequences with lengths >100 bp and gap 'N' less than 5% constituted the raw transposable element (TE) library. A custom library (a combination of Repbase and our de novo TE library which was processed by uclust to yield a non-redundant library) was supplied to RepeatMasker for DNA-level repeat identification. A total of 1,832,346 repeat sequences were identified, with a cumulative length of 1,117,418,804 bp, accounting for 70.02% of the genome. Among them, the most abundant are LTR elements, with a total of 667,992 elements spanning 602,850,190 bp, accounting for 37.78% of the entire genome (Table S10).

Structural annotation of the genome incorporates ab initio prediction, homology-based prediction and RNA-Seq assisted prediction, was used to annotate gene models. Sequences of homologous proteins were

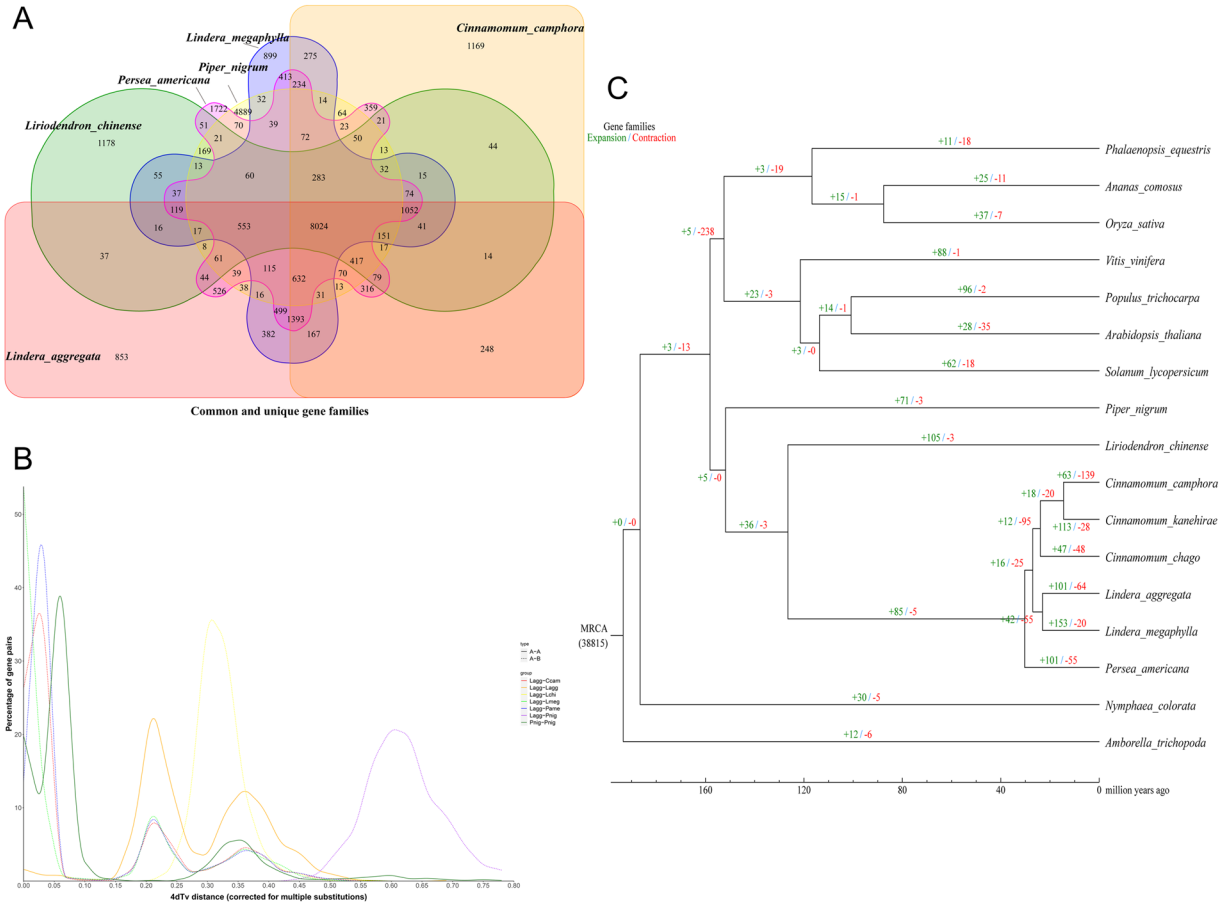

Fig. 3 Gene families and evolutionary analysis. (**A**) Unique and common gene clusters between *L. aggregate* and five related species. (**B**) Detection of whole genome duplication (WGD) events based on fourfold degenerate synonymous site (4DTv). (**C**) Analysis of gene family expansion and contraction among *L. aggregate* and 16 other species.

downloaded from Ensembl and NCBI (including *Populus trichocarpa*, *Arabidopsis thaliana*, *Cinnamomum kanehirae*, *C. chago*, *C. camphora*, *Persea americana*, *Lindera megaphylla*). Protein sequences were aligned to the genome using TblastN v2.2.26[24], and then the matching proteins were aligned to the homologous genome sequences for accurate spliced alignments with GeneWise v2.4.1 software[25] which was used to predict gene structure contained in each protein region. For gene predication based on Ab initio, Augustus v3.5[26] and SNAP v2013.11.29[27] were used in our automated gene prediction pipeline. Transcriptome reads assemblies were generated with Trinity v2.8.5[28] for the genome annotation. To optimize the genome annotation, the RNA-Seq reads from different tissues which were aligned to genome fasta using Hisat v2.2.1[29] with default parameters to identify exons region and splice positions. The alignment results were then used as input for Stringtie v2.2.1[30] with default parameters for genome-based transcript assembly. The non-redundant reference gene set was generated by merging genes predicted by three methods with EvidenceModeler v1.1.1[31] using PASA[32] terminal exon support and including masked transposable elements as input into gene prediction (Table S11).

Gene functions were assigned according to the best match by aligning the protein sequences to the Swiss-Prot using Blastp v2.2.26[24]. The motifs and domains were annotated using InterProScan v5.39[33] by searching against publicly available databases, including ProDom, PRINTS, Pfam, SMRT, PANTHER and PROSITE. The Gene Ontology (GO) IDs for each gene were assigned according to the corresponding InterPro entry. We predicted the proteins function by transferring annotations from the closest BLAST hit in the SwissProt and NR database via DIAMOND v0.8.22[34]. We also mapped gene set to a KEGG pathway and identified the best match for each gene. Finally, in the above-mentioned at least one database, 32,268 genes were functionally annotated, accounting for 96.95% of the all genes (Table S12).

The tRNAs were predicted using the program tRNAscan-SE v1.4[35]. For rRNAs are highly conserved, we choose relative species' rRNA sequence as references, predict rRNA sequences using Blast. Other ncRNAs, including miRNAs, snRNAs were identified by searching against the Rfam database[36] with default parameters using the Infernal v1.1.5 software (Table S13).

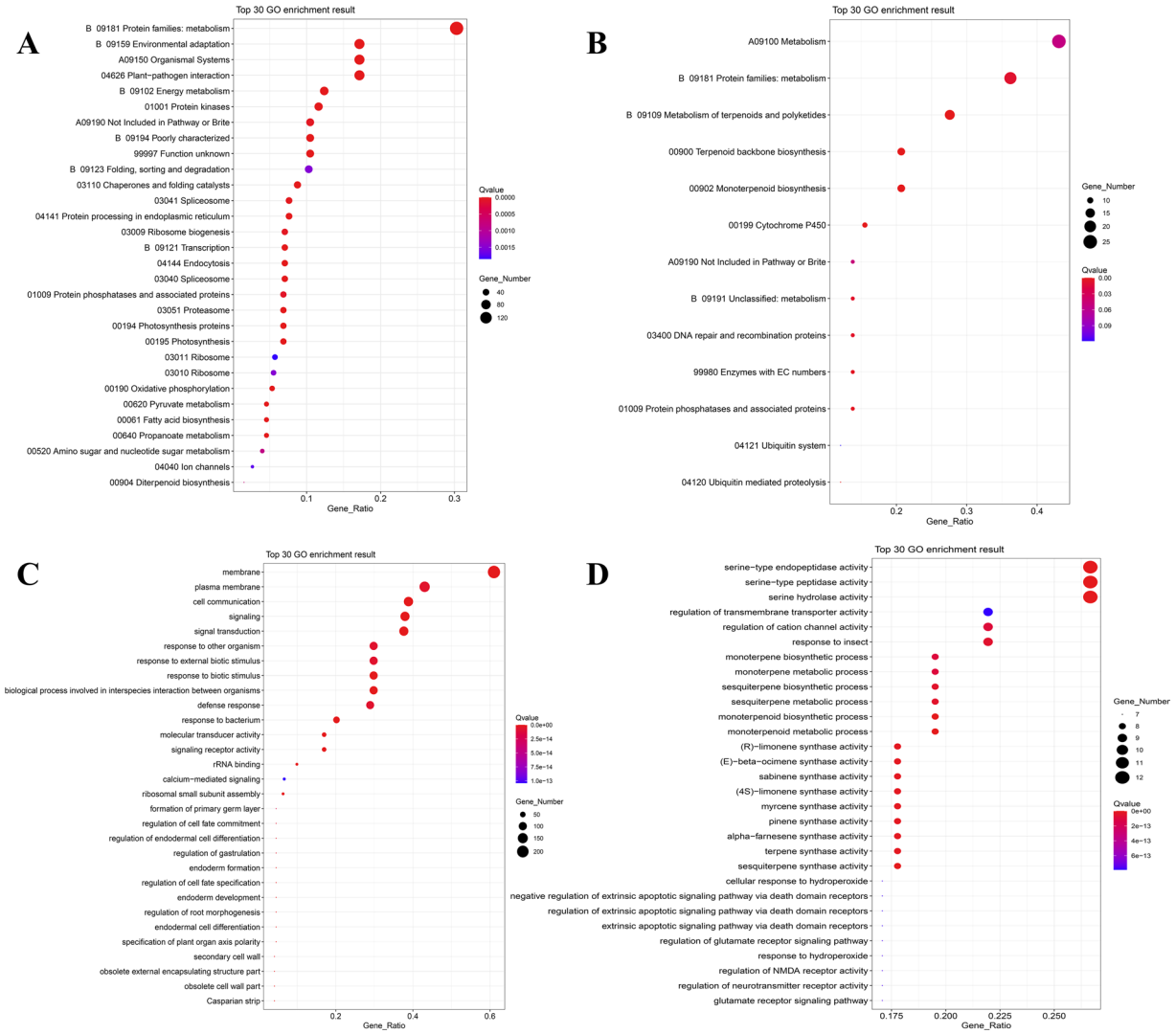

Fig. 4 Enrichment analysis of KEGG and GO pathways. KEGG enrichment analysis of expansion genes (**A**) and contraction genes (**B**) in *L. aggregate*. GO enrichment analysis of expansion genes (**C**) and contraction genes (**D**). The GO terms and KEGG pathways with a p-value < 0.05 were defined as statistically significant.

Data Records

The relevant data reported in this manuscript has been stored in GenBank, with BioProject accession PRJNA1144506. The PacBio HiFi reads, Hi-C reads, WGS data, and RNA-Seq data have been saved in the SRA of NCBI with accession SRR31617073[37], SRR31617077[38], SRR31617074[39], SRR31617078[40], SRR31617072[41], SRR31617071[42], SRR31617069[43], SRR31617067[44], SRR31617068[45], SRR31617070[46], SRR31617076[47], SRR31617075[48]. Data for the final chromosome assembly is stored in GeneBank accession JBLYEH000000000[49]. Genome sequence and annotation data can also be found in Figshare[50]. The accession numbers of mitochondrial and chloroplast genomes of *Lindera aggregata* in Gene Bank are PP848112[51], PP848113[52] and PP199190[53].

Technical Validation

Genome assembly quality assessment. The final assembled genome was about 1.59 Gb, similar to K-mer prediction (Table S6; Figure S1). The integrity of the assembly sequence was assessed using BUSCO V5.4.5[54] tool. The proportion of complete core genes (including single copy and multiple copy genes) was 97.3%, and the proportion of missing genes was 1.8%, indicating that the assembled genome has good integrity (Fig. 2; Table S7).

In order to evaluate the accuracy of assembly, small fragment library reads were selected and compared to the assembled genome using BWA v5.4.5[55] software. The alignment rate of reads, the extent of coverage of the genome and the distribution of depth were calculated to evaluate the integrity of assembly and the uniformity of sequencing. The results show that the comparison rate of short reads to the genome was about 98.15%, and the genome coverage rate was approximately 99.92%, indicating that the reads and the assembled genome have good consistency (Table S8).

The Merqury v1.4.1[56] software was then used to evaluate the quality of the genome based on K-mer. The Qv of the assembled genome was approximately 40.13, indicating that the assembly accuracy exceeded 99.99% (Table S9).

The strength of the interaction signals around the diagonal of the genome-wide Hi-C heat map was stronger than the off-diagonal signals (Fig. 1D), indicating the high quality of genome assembly at the chromosomal level.

Evaluation of the gene annotation. The annotated proteins were evaluated via BUSCO with the lineage dataset embryophyta_odb10. The assessment results showed that a total of 88.9% of complete BUSCOs (84.4% complete and single-copy BUSCOs and 4.5% complete Duplicated BUSCOs) were annotated in the gene set of *L. aggregate*, and only 5.4% of the genes were missing BUSCOs, indicating that the annotation results were high quality (Table S14).

Gene family analysis. Using gene family cluster analysis, the genomes of L. aggregate and 16 other sequenced plants (Persea americana, Cinnamomum camphora, Cinnamomum kanehirae, Cinnamomum chago, Lindera megaphylla, Liriodendron chinense, Piper nigrum, Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Solanum lycopersicum, Oryza sativa, Ananas comosus, Phalaenopsis equestris, Amborella trichopoda, Nymphaea colorata) were analyzed. The results showed that a total of 38,831 gene families were clustered among 17 species; there were 5527 shared gene families, of which 381 single-copy gene families were shared by each species (Figure S3). Further comparison with 5 related species (C. camphora, L. chinense, L. megaphylla, P. americana, P. nigrum) found that 6 species had a total of 8024 gene clusters. In addition, compared with the other five genomes, we found 3092 unique gene families in the Lindera genome (Fig. 3A).

Comparative genome analysis. Based on a comparison of fourfold degenerate synonymous site (4DTv) in common linear blocks among *L. aggregate* and its related species, it was revealed that whole genome duplication (WGD) events shared by *L. aggregate* and other Lauraceae plants (Fig. 3B). Furthermore, 16 other sequenced plant genomes were used for the expansion and contraction of gene families, showing that 101 gene families in *L. aggregate* had significantly expanded and 64 gene families had significantly contracted. The results indicated that the gene familys of *L. aggregate* had mainly experienced expansion during the adaptive evolution process (Fig. 3C).

The expanded gene families were found to be primarily enriched in KEGG pathways related to metabolism, environmental adaptation, organismal systems, plant-pathogen interaction, energy metabolism (Fig. 4A). In terms of GO functions, enrichment was detected in membrane, signal transduction, response to external biotic stimulus, response to biotic stimulus, biological process involved in interspecies interaction between organisms (Fig. 4C). On the other hand, the contracted gene families were mainly enriched in the GO function of serine-type endopeptidase activity, regulation of transmembrane transporter activity, monoterpene biosynthetic process, sesquiterpene biosynthetic process, monoterpenoid biosynthetic process (Fig. 4D). In addition, in terms of the KEGG pathways, enrichments were observed in metabolism, metabolism of terpenoids and polyketides, terpenoid backbone biosynthesis, monoterpenoid biosynthesis and cytochrome P450 (Fig. 4B). As with most plants in the Lauraceae family[57–59], the terpenoid biosynthesis pathway is the key route for the synthesis of active components in *L. aggregata*.

Code availability

All commands used were executed in accordance with the manuals or protocols of the tools used in this study. The software and tools used are publicly accessible, and the version and parameters are specified in the Methods section. If detailed parameters are not mentioned, default parameters are used. No custom codes were used in this study.

Received: 13 December 2024; Accepted: 24 March 2025;
Published online: 03 April 2025

References

1. Cao, Y. *et al.* The genus *Lindera*: a source of structurally diverse molecules having pharmacological significance. *Phytochemistry Reviews* **15**, 869–906, https://doi.org/10.1007/s11101-015-9432-2 (2016).
2. Tao, Y., Deng, Y. & Wang, P. Traditional uses, phytochemistry, pharmacology, processing methods and quality control of *Lindera aggregata* (Sims) Kosterm: A critical review. *Journal of Ethnopharmacology* **318**, 116954, https://doi.org/10.1016/j.jep.2023.116954 (2024).
3. Lv, Y. *et al.* A review on the chemical constituents and pharmacological efficacies of *Lindera aggregata* (Sims) Kosterm. *Frontiers In Nutrition* **9**, 1071276, https://doi.org/10.3389/fnut.2022.1071276 (2022).
4. Duong, T.-H. *et al.* Atypical Lindenane-Type Sesquiterpenes from Lindera myrrha. *Molecules* **25**, 1830, https://doi.org/10.3390/molecules25081830 (2020).
5. Shi, Y., Chen, Z., Jiang, J., Li, X. & Zeng, W. Comparative Analysis of Chloroplast Genomes of "Tiantai Wu-Yao" (*Lindera aggregata*) and Taxa of the Same Genus and Different Genera. *Genes (Basel)* **15**, 263, https://doi.org/10.3390/genes15030263 (2024).
6. Peng, X. *et al.* Integrated analysis of the transcriptome, metabolome and analgesic effect provide insight into potential applications of different parts of Lindera aggregata. *Food Research International* **138**, 109799, https://doi.org/10.1016/j.foodres.2020.109799 (2020).
7. Chen, K. W. *et al.* Effects of external qigong therapy on osteoarthritis of the knee. A randomized controlled trial. *Clinical Rheumatology* **27**, 1497–1505, https://doi.org/10.1007/s10067-008-0955-4 (2008).
8. Lee, H. J. *et al.* Turo (qi dance) training attenuates psychological symptoms and sympathetic activation induced by mental stress in healthy women. *Evidence-based Complementary And Alternative Medicine* **6**, 399–405, https://doi.org/10.1093/ecam/nem120 (2009).
9. Wen, S.-S. *et al.* Characterization and quantification of the phytochemical constituents and anti-inflammatory properties of Lindera aggregata. *RSC Advances* **14**, 36101–36114, https://doi.org/10.1039/D4RA05643D (2024).

10. Huang, Q., Liu, K., Qin, L. & Zhu, B. Lindera aggregata (Sims) Kosterm: a systematic review of its traditional applications, phytochemical and pharmacological properties, and quality control. *Medicinal Plant Biology* **2**, 11, https://doi.org/10.48130/MPB-2023-0011 (2023).
11. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15, https://worldveg.tind.io/record/33886 (1987).
12. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
13. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
14. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
15. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
16. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
17. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
18. Jin, J.-J. *et al*. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* **21**, 241, https://doi.org/10.1186/s13059-020-02154-5 (2020).
19. Bi, C. *et al*. PMAT: an efficient plant mitogenome assembly toolkit using low-coverage HiFi sequencing data. *Horticulture Research* **11**, uhae023, https://doi.org/10.1093/hr/uhae023 (2024).
20. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
21. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 1–6, https://doi.org/10.1186/s13100-015-0041-9 (2015).
22. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4.10. 11–14.10. 14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
23. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
24. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**, W20–W25, https://doi.org/10.1093/nar/gkh435 (2004).
25. Birney, E. & Durbin, R. Using GeneWise in the Drosophila annotation experiment. *Genome Research* **10**, 547–548, https://doi.org/10.1101/gr.10.4.547 (2000).
26. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**, W465–W467, https://doi.org/10.1093/nar/gki458 (2005).
27. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 1–9, https://doi.org/10.1186/1471-2105-5-59 (2004).
28. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
29. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
30. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
31. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
32. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
33. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
34. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
35. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
36. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **49**, D192–D200, https://doi.org/10.1093/nar/gkaa1047 (2020).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617073 (2024).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617077 (2024).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617074 (2024).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617078 (2024).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617072 (2024).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617071 (2024).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617069 (2024).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617067 (2024).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617068 (2024).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617070 (2024).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617076 (2024).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31617075 (2024).
49. *NCBI Assembly* https://identifiers.org/insdc:JBLYEH000000000 (2024).
50. Shi, Y. *et al*. Chromosome-level genome assembly of the traditional medicinal plant *Lindera aggregata*. *Figshare* https://doi.org/10.6084/m9.figshare.28007183 (2024).
51. Shi, Y. *et al*. *Lindera aggregata* chloroplast, complete genome. *GenBank* https://identifiers.org/ncbi/insdc:PP199190 (2024).
52. Shi, Y. *et al*. *Lindera aggregata* chromosome 1 mitochondrion, complete sequence. *GenBank* https://identifiers.org/ncbi/insdc:PP848112 (2024).
53. Shi, Y. *et al*. *Lindera aggregata* chromosome 2 mitochondrion, complete sequence. *GenBank* https://identifiers.org/ncbi/insdc:PP848113 (2024).
54. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
56. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).

57. Xiong, B. *et al*. Genome of Lindera glauca provides insights into the evolution of biosynthesis genes for aromatic compounds. *Iscience* **25**, 104761, https://doi.org/10.1016/j.isci.2022.104761 (2022).
58. Han, X. *et al*. The chromosome-scale genome of Phoebe bournei reveals contrasting fates of terpene synthase (TPS)-a and TPS-b subfamilies. *Plant Communications* **3**, 100410, https://doi.org/10.1016/j.xplc.2022.100410 (2022).
59. Chen, Y.-C. *et al*. The Litsea genome and the evolution of the laurel family. *Nature Communications* **11**, 1675, https://doi.org/10.1038/s41467-020-15493-5 (2020).

Acknowledgements

This work was funded by the Basic Public Welfare Research Project of Zhejiang Province (LGN22C020001), Taizhou 500 talent program (Z2024136) and Key Scientific and Technological Grant of Zhejiang for Breeding New Agricultural Varieties (2021C02074).

Author contributions

Y.J.S., Z.C., J.X.G., J.Y.J., Q.F.L., Y.L.L., W.F.Y. and W.Z. conceived and performed the original research project. Y.J.S., J.Y.J., Q.F.L., Y.L.L. and W.F.Y. collected samples and performed the experiments. Y.J.S., Z.C. and W.Z. designed the experiments and analyzed the data. Y.J.S. refined the project and wrote the manuscript with contributions from all authors. Z.C. and W.Z. supervised the experiments and revised the writing. Z.C., J.X.G. and W.Z. obtained the funding for the research project. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04891-3.

Correspondence and requests for materials should be addressed to W.Z.