Exploration of the Species-specific Dna Markers Based on the Complete Chloroplast Genome for Discriminating Curcuma Comosa Roxb. From Curcuma Latifolia Roscoe and Other Related Species

Bussarin Wachananawat
Chulalongkorn University

Bobby Lim-Ho Kong
Li Dak Sum Yip Yio Chin R & D Centre for Chinese Medicine and Institute of Chinese Medicine, The Chinese University of Hong Kong

Pang-Chui Shaw
Li Dak Sum Yip Yio Chin R & D Centre for Chinese Medicine and Institute of Chinese Medicine, The Chinese University of Hong Kong

Bhanubong Bongcheewin
Mahidol University

Sunisa Sangvirotjanapat
Mahidol University

Pinidphon Prombutara
Chulalongkorn University

Natapol Pornputtapong
Chulalongkorn University

Suchada Sukrong (✉ suchada.su@chula.ac.th)
Chulalongkorn University

Research Article

Keywords: Curcuma, nucleotide barcodes, rhizomatous, herbs

Posted Date: December 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1141755/v1

Abstract

Members of the *Curcuma* genus are among the most commonly used rhizomatous herbs worldwide. There are two species of *Curcuma* referred to as “Wan Chak Motluk” in Thai, *C. comosa* Roxb. and *C. latifolia* Roscoe, and their herbal materials are often confused. *C. comosa* is widely used as a traditional herbal remedy for its phytoestrogenic activity, but its morphology is highly similar to that of *C. latifolia*, which contains a compound that causes hepatotoxicity. In this study, the complete chloroplast (cp) genomes of these species were determined for the first time using Illumina sequencing. Our results showed that their cp genomes were 162,272 bp (*C. comosa*) and 162,289 bp (*C. latifolia*) in length. A total of 133 unique genes were identified, including 87 protein-coding genes, 38 tRNA genes and 8 rRNA genes. Comparative analyses with other species of *Curcuma* indicated high similarity in gene content and structural organization. The analyses also reveal variable hotspots in the genomes at *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V that can serve as species-specific nucleotide barcodes. Indeed, mislabeling of these two species among samples sold at market was detected using these species-specific markers, indicating that cp genomes can provide more information for better elucidating and improving discriminatory power for species authentication.

Introduction

Curcuma, a rhizomatous genus, belongs to the Zingiberaceae, a family of approximately 120 species distributed throughout tropical Asia, Australia, and South Pacific Islands[5]. Many species of *Curcuma*, such as *C. longa* L., *C. aromatica* Salisb., and *C. zedoaria* (Christm.) Roscoe have been used in natural food additives, cosmetics and traditional medicine[36]. Apart from those species, *C. comosa*, or “Wan Chak Motluk” in Thai, is an economically important species that is widely used as a traditional herbal remedy for estrogenic hormone deficits in women and has a protective effect on postmenopausal osteoporosis[33,34,44]. Three morphologically similar species of *Curcuma*, namely, *C. latifolia*, *C. elata*, and *C. xanthorrhiza*, share the common name “Wan Chak Motluk” and are often confused for *C. comosa* in herbal markets. *C. latifolia* is the species most often incorrectly substituted for *C. comosa* because of their similarity in appearance[39]. However, *C. comosa* and *C. latifolia* have different active compounds and activities. *C. comosa* contains numerous diarylheptanoids as major active compounds. It possesses phytoestrogenic properties and be used as a supplement for treating uterine and ovarian abnormalities[3,18,41,45,46]. *C. latifolia* contains a sesquiterpenoid compound, zederone, that can cause hepatotoxicity[18,32]. Therefore, the misidentification of *C. latifolia* as *C. comosa* has become a major concern for consumers and the herbal industry. The discrimination of *C. comosa* from other related species, including *C. latifolia*, is important for consumer safety.

Many identification methods have been used to authenticate *Curcuma* spp. The traditional method is to use taxonomic keys based on morphological data. In particular, the inflorescence is suitable for the identification of these *Curcuma* spp. However, inflorescences are not always available in a complete form due to the short-lived and highly seasonal flowering of these species. The limited inflorescence availability makes it extremely difficult to identify *Curcuma* spp.[27,36]. Furthermore, the large variation in rhizome morphology among Zingiberaceae can lead to confusion in herbal usage (Fig. 1). The Thai Herbal Pharmacopoeia (THP), an official national standard compendium, provides standard quality control for many herbal drugs marketed in Thailand; however, there are no data available on *C. comosa* and *C. latifolia*. Only morphological characteristics are used to discriminate *C. comosa* from *C. latifolia*, but this method is still unclear. Thus, other methods for differentiating these plants are required, such as chemical profiling, molecular cytogenetics, and molecular marker analysis. However, chemical patterns also have some limitations. For example, chemical contents can be altered due to cultivation or weather conditions and harvesting time[9]. A few molecular cytogenetics studies have reported that *C. comosa* cultivars have chromosome numbers 2n = 63 and 42[16,30], whereas *C. latifolia* consists of two cultivars with 2n = 63 and 84[39]. The number of chromosomes has been changed by mutation, chromosome doubling, and hybridization. Recently, many DNA molecular technologies have been developed as tools for the differentiation of various plants in *Curcuma* spp., including random amplified polymorphic DNA (RAPD)[2], inter simple sequence repeat (ISSR)[38], and amplified fragment length polymorphism (AFLP)[37] techniques. DNA barcoding is another molecular technique that can be used for species identification. Four DNA barcodes based on chloroplast DNA regions (*mat*K, *rbc*L, and *trn*H-*psb*A) and a nuclear DNA region (*ITS*) have been established for the identification of *Curcuma* spp. Other chloroplast DNA regions, such as *rpo*C1, *rpo*B, *rps*36-*rps*8, *ndh*J, *trn*L-F, *acc*D, and *trn*S-*trn*fM, have been evaluated as barcode loci in Zingiberaceae[29,43]. Although traditional barcodes have been widely studied, they offer a limited ability to distinguish closely related species[14,19].

Because of the lack of adequate variations in DNA barcode sequences, a new method is needed to identify closely related plant species. Chloroplast genome sequencing is one such tool and has recently been shown to successfully discriminate closely related species in the genera *Boesenbergia*, *Curcuma*, *Kaempferia*, and *Pyrgophyllum*[25,26].

Therefore, this study has identified species-specific DNA markers to allow the discrimination of *C. comosa* from other related species, including *C. latifolia*, based on the cp genome. The developed species-specific DNA markers were then utilized for testing crude drugs sold in herbal markets. Additionally, this study provides reliable and high-quality chloroplast genome resources for future research in Zingiberaceae.

Results

Chloroplast genome organization and features of *C. comosa* and *C. latifolia*. Illumina high-throughput sequencing was used to obtain the cp genome sequences of *C. comosa* and *C. latifolia*. The *C. comosa* and *C. latifolia* cp genomes were 162,272 bp and 162,289 bp in length, respectively (Fig. 2). Both cp genomes had typical quadripartite structures consisting of a large single-copy (LSC) region of 87,074 bp in *C. comosa* and 87,089 bp in *C. latifolia*, a small single-copy (SSC) region of 15,698 bp in *C. comosa* and 15,700 bp in *C. latifolia*, and two inverted repeat (IR) regions of 29,750 bp in both species. The GC contents of both cp genomes were identical, at 34% (LSC), 29.7% (SSC), 41.2% (IR), and 36.2% (total). In addition, the cp genomes of the two species were predicted to encode 133 genes, including 87 protein-coding genes, 38 tRNA genes and 8 rRNA genes (Table 1). These genes were classified into 4 major groups according to their functions, including self-replication (4 ribosomal RNA genes, 30 transfer RNA genes, 12 small ribosomal subunit genes, 9 large ribosomal subunit genes, and 4 DNA-dependent RNA polymerase genes), photosynthesis (5 photosystem I genes, 14 photosystem II genes, 6 cytochrome b/f complex genes, 6 ATP synthase genes, 1 ATP-dependent protease gene, 1 Rubisco large subunit gene, and 11 NADH dehydrogenase genes), other (maturase, envelope membrane protein, acetyl-CoA-carboxylase, c-type cytochrome synthesis, and translation initiation factor), and unknown (4 genes) (Table 2). Intragenic regions were found in 18 genes. Of these, 13 genes (*rps*16*, rpo*C1*, atp*F*, pet*B*, pet*D*, rps*12*, rpl*16*, ycf*3*, clp*P*, trn*K-UUU*, trn*L-UAA*, trn*V-UAC*,* and *trn*G-UCC), 1 gene (*ndh*A), and 4 genes (*ndh*B*, rpl*2*, trn*I-GAU*,* and *trn*A-UGC) were in the LSC, SSC and IR regions, respectively (Fig. 2; Table 2).

Codon usage. The relative synonymous codon usage (RSCU) of ten species belonging to five different genera in the family Zingiberaceae (Table S1), including *C. comosa* and *C. latifolia*, was calculated. There was no bias in the usage of the start codons methionine (AUG) and tryptophan (UGG) (RSCU = 1). The 87 protein-coding genes contained approximately 28,393 codons. UUA-encoded leucine had the highest RSCU, at approximately 1.94, and GCG-encoded alanine had the lowest RSCU, at approximately 0.39. All preferred synonymous codons with A or U at the third position showed higher bias (RSCU >1) than those with G or C (Table S2).

Repeat structure analysis. The cp genome sequences of the plants in Zingiberaceae (Table S1) were retrieved for SSR and long repeat analysis using MISA software and the REPuter program. A total of 78-121 SSRs were found in the cp genomes of the ten species (Fig. 3; Table S3). Among the different types of SSRs, mononucleotide repeats were the most abundant, accounting for 27-58 loci, followed by dinucleotide (32-34 loci), tetranucleotide (17-21 loci), trinucleotide (3-8 loci), pentanucleotide (1-4 loci), and hexanucleotide (0-2 loci) repeats (Fig 3A; Table S3). Mononucleotide SSRs were especially rich in A/T repeats (239-280 loci) (Fig 3B; Table S3). The SSR repeats were mainly distributed in the LSC regions (51-79 loci), while only a small portion were located in the SSC regions (13-22 loci) and IR regions (5-8 loci) (Fig. 3C; Table S3). The long repeat analysis identified a total of 39-79 long repeat sequence types (Fig. 4; Table S4). Among the different types of long repeats, forward repeats (9-28 loci) were the most abundant, followed by palindromic (8-28 loci), reverse (4-16 loci), and complement (1-10 loci) repeats (Fig. 4A; Table S4). Repeat lengths of 30-39 bp were the most abundant among the ten cp genomes used in this study (Fig. 4B; Table S4).

Highly variable sequences in noncoding regions of the cp genome of *C. comosa* and *C. latifolia*. To compare the sequence divergences of *C. comosa* and *C. latifolia*, the cp genome sequences of the 10 species in Zingiberaceae (Table S1) were included for comparison using the mVISTA program, and *C. comosa* was used as the reference. Overall, the coding regions were more conserved than the noncoding regions among the 10 species of Zingiberaceae; however, *rpo*C2*, rpo*B*, ycf*1*, ycf*2*,* and *ndh*F exhibited some degree of variation. The two IR regions were less divergent than the LSC and SSC regions. In contrast, high levels of divergence were found in the intergenic regions of *trn*K*-rps*16*, rpo*B*-trn*C*, rps*4*-trn*T*, trn*T*-trn*L*, ndh*C*-trn*V*,* and *ndh*F*-rpl*32 (Fig.

5). In addition to nucleotide divergence, the expansion and contraction of the border regions were also analyzed for the 10 species of Zingiberaceae (Fig. 6; Table S1). The four junctions, LSC/IRa, LSC/IRb, SSC/IRa and SSC/IRb, were found to be almost the same (29,642 bp to 29,797 bp). The *rpl*22-*rps*19 genes were located at the boundary of the LSC/IRb region in each cp genome. The *rpl*22 gene was located on the left side of the LSC/IRb boundary, at a distance of 21 bp to 48 bp. The *rps*19 gene was located on the right side of the LSC/IRb boundary, at a distance of 129 bp to 148 bp. The *ycf*1-*ndh*F genes were located at the IRb/SSC boundary. The IRb/SSC junction was located in the *ycf*1 region and extended a length of 7 bp to 205 bp into the SSC region. The *ndh*F gene was located on the right side of the IRb/SSC boundary, at a distance of 8 bp to 218 bp. The SSC/IRa junctions in the cp genomes were embedded in the *ycf*1 genes, with the distance of 3,705 bp to 3,899 bp in the IRa region. The *rps*19-*psb*A genes were located at the boundary of the IRa/LSC region. The *rps*19 gene was located on the left side of the IRa/LSC boundary, at a distance of 129 bp to 148 bp, while *psb*A was located on the right side of the IRa/LSC boundary, at a distance of 109 bp to 125 bp.

***ndh*A, *Trn*T-*trn*L, and *ndh*C-*trn*V are DNA signature sites for the development of species-specific markers.** The cp genome sequences belonging to 33 species of Zingiberaceae (Table S5) were analyzed for species-specific DNA markers. As expected, the sliding window analysis showed the most variation in the LSC and SSC regions but lower variation in the IR regions (Fig. 7). The average value of nucleotide diversity (Pi) was 0.0096 among the 33 Zingiberaceae species (Table S6). Mutational hotspots were found in 6 genes, *rps*16-*trn*Q*, ycf*1*, ndh*A*, ndh*I*, ndh*D*,* and *RF*19; these sites exhibited remarkable Pi values, higher than 0.03 (Fig. 7A). In addition, the average value of nucleotide diversity (Pi) among 20 species in *Curcuma* was 0.0018 (Table S6), and there were 3 mutational hotspots in *rps16-trnQ, petN-psbM,* and *ndhA* that exhibited Pi values higher than 0.01 (Fig. 7B). Additionally, there were 6 SNPs and 41 indels in the cp genomes of *C. comosa* and *C. latifolia* (Table S7). When comparing the cp genomes among the 10 selected species of Zingiberaceae (Table S1), SNP/indel variation sites were found in the *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V regions, with 1 SNP, a 6 bp insertion, and a 2 bp deletion, respectively (Figure S1).

Validation of the species-specific DNA markers in crude "Wan Chak Motluk" sold at market. We analyzed 19 samples of crude drugs claiming to be "Wan Chak Motluk", comprising 14 samples of *C. comosa* and 5 samples of *C. latifolia*, represented as CD-01 to CD-19 (Table S8). All samples were purchased from various herbal markets. PCR amplification of *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V with our developed species-specific primers yielded products of 330 bp, 264 bp, and 370 bp in length, respectively (Table S9). Of the 14 samples claiming to be *C. comosa*, 5 samples (CD-01, CD-07, CD-16, CD-18, and CD-19) were confirmed as *C. comosa,* while 8 samples (CD-02, CD-03, CD-04, CD-05, CD-06, CD-08, CD-09 and CD-17) were identified as *C. latifolia* (Table 3), and one sample (CD-10) was neither *C. comosa* nor *C. latifolia* (Table 3). Examination of 5 samples (CD-11, CD-12, CD-13, CD-14, and CD-15) claiming to be *C. latifolia* revealed that only 3 samples (CD-12, CD-13, and CD-14) were *C. latifolia,* and the remaining 2 samples (CD-11 and CD-15) were neither *C. latifolia* nor *C. comosa* (Table 3).

Phylogeny construction with the cp genome sequences of *C. comosa* and *C. latifolia*. To examine the phylogenetic positions of the *C. comosa* and *C. latifolia* species and their relationships within Zingiberales (Table S10), neighbor-joining (NJ) phylogenetic analyses were performed using 40 cp genomes from 40 species belonging to 6 families of Zingiberales. In this analysis, six families in Zingiberales were divided into two clades with 100% bootstrap support (BS) values. One clade was composed of five families, including Musaceae, Heliconiaceae, Strelitziaceae, Cannaceae, and Costaceae, while the other clade included only Zingiberaceae. The clade containing Zingiberaceae was divided into 2 groups. The first group included four genera (*Wurfbainia, Amomum, Lanxangia,* and *Alpinia*) (BS =100%), and the second group included five genera (*Curcuma, Stahlianthus, Hedychium, Kaempferia,* and *Zingiber*). The second group was further divided into 4 subgroups (BS = 72-100%). Subgroup II was the most complex, with 17 species, including the species of interest, *C. comosa* and *C. latifolia,* on the same branch as *C. elata* and *C. aromatica* (BS = 100%). (Fig. 8).

Discussion

The accurate discrimination of herbal material using only morphological characteristics is difficult. DNA barcoding has become a conventional technology used for such discrimination. Although DNA barcoding technology has been developed significantly, no barcode has yet achieved the goal of reliable identification of all plant species. In fact, the identification of species with close genetic relationships continues to pose great challenges[10]. In this study, DNA regions including the *rbc*L, *mat*K, *psb*A-*trn*H spacer and *ITS*2 of two authentic *C. comosa* and *C. latifolia* were established; their accession numbers have been submitted to GenBank (Table S11). Unfortunately, there were no variations observed in any of these core barcode regions. This result is consistent with previous studies showing low variation in the nucleotide sequences of these core barcode regions within closely related species, including species within *Curcuma*[6], *Chrysanthemum*[14], *Cymbidium*[48] and *Ligularia*[7]. Therefore, we developed specific markers for the discrimination of *C. comosa* from *C. latifolia* and other related species belonging to Zingiberaceae based on mutation hotspot regions identified in cp genomes. In previous reports, closely related plant species in *Hedyotis*[49], *Gentiana*[51], and *Fritillaria*[47] were successfully identified at the species level based on divergent sequences at hotspots in the cp genome. The cp genomes of many *Curcuma* spp. have been studied, and divergent hotspots have been proposed as authentication markers,[26] but these prior efforts excluded *C. comosa* and *C. latifolia*. In this research, no significant variation in the numbers of total genes, protein-coding genes, tRNAs or rRNAs was shown between *C. comosa* and *C. latifolia*. Similar findings were observed in other *Curcuma* spp.[13,26] and other plant species in Zingiberaceae[8,23,25]. Codon usage compares the frequencies of each three-nucleotide sequence that codes for a particular amino acid[4]. Codons are used in transmitting genetic information and serve as the building blocks of proteins[24]. Codon usage is a factor that has shaped the evolution of cp genomes due to biases in mutation, and it varies across species[22]. The results of the codon usage analysis showed that codon usage bias was low in the cp genomes of *C. comosa* and *C. latifolia*, indicating their similar evolutionary paths. Repeat structure plays an important role in genomic rearrangement, recombination and sequence divergence in plastomes[40]. However, we did not find significant variations in repeat distribution, especially in *Curcuma* genera (Fig. 4). Despite the fact that the organization of the cp genome is highly similar among these Zingiberaceae species, we discovered several regions with interspecific polymorphisms, mostly located within intergenic regions and at the LSC, SSC and IR regional boundaries of the cp genomes (Fig. 5-7). Among the 20 *Curcuma* cp genomes analyzed, including *C. comosa*, *C. latifolia* and *Curcuma* spp. sequences retrieved from the NCBI database, three divergent hotspots, *rps*16-*trn*Q, *pet*N-*psb*M and *ndh*A, were found. This finding agreed well with a previously published report on the cp genome of *Curcuma* spp. in Zingiberales[26]. Based on our alignment of the cp genome, *C. comosa* has one unique sequence at *ndh*A. However, the sequences of *rps*16-*trn*Q and *pet*N-*psb*M could not discriminate *C. comosa* from *C. latifolia*. Therefore, SSRs, indels and SNPs were examined for use as molecular markers. After searching, indel-variable loci at *trn*T-*trn*L and *ndh*C-*trn*V were discovered and used for differentiation of *C. comosa* from other related species, including *C. latifolia*. In our study, three regions, *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V, were successfully defined as species-specific DNA markers for *C. comosa*. In previous publications, these same three regions were recommended for the identification of plants in the genera *Dendrobium*[50], *Actaea*[31], and *Entandrophragma*[28]. Then, nineteen crude drug samples claimed to be “Wan Chak Motluk” (*C. comosa* or *C. latifolia*) purchased from different herbal markets (Table S8) were examined using our developed species-specific DNA markers, including *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V regions (Figure S1; Table S9). Surprisingly, only 5 out of 19 samples were confirmed as *C. comosa*, while 11 samples were identified as *C. latifolia*. Three samples were neither *C. comosa* nor *C. latifolia*. Our crude drug identification indicated that these two species are widely mislabeled in herbal markets. Thus, the Thai Food and Drug Administration (Thai FDA) must become involved and be notified about the quality of herbal materials as well as consumer safety. Therefore, the species-specific DNA marker developed through cp genome sequencing could be used as one tool for authenticating the origin of these plant species. This marker can help to resolve the limitations of core DNA barcoding in discriminating closely related plant species. With the genetic variation discovered, we also analyzed the phylogenetic relationships of the 20 species of *Curcuma* and other species of Zingiberales using cp genomes. The phylogenetic analysis revealed that all of the species of *Curcuma* formed a complex monophyletic clade with high bootstrap support values. As expected, *C. comosa* and *C. latifolia* exhibited a very close relationship in our phylogenetic analysis.

Conclusion

Cp genome sequences represent a valuable genomics resource for exploring diversity in any plant family. To the best of our knowledge, this is the first study of the cp genome sequences of *C. comosa* and *C. latifolia*, plants in Zingiberaceae. Within these cp genomes, species-specific nucleotide sequences that can be used as DNA markers for discriminating *C. comosa* from other related species, including *C. latifolia*, were discovered in the *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V regions. Species-specific DNA markers can be used for species authentication when the origin or morphological characteristics of plants or herbal materials are not available or sufficient for evaluation. The constructed phylogenetic tree of *C. comosa* and *C. latifolia* and related species provides a deeper understanding of the relationships among plants in the genus *Curcuma* and the family Zingiberaceae and can be exploited as a reliable resource for further Zingiberaceae research. To ensure the safe and effective use of herbal or raw materials, species identification and quality control steps should be performed before their release onto the markets.

Materials And Methods

Plant materials and crude drugs. Two cultivated authentic species, *C. comosa* Roxb. and *C. latifolia* Roscoe, were collected from Zingiberaceae Collection, Sireeruckhachat Nature Learning Park, Nakhon Pathom, Thailand and identified by taxonomist Dr. Bhanubong Bongcheewin at Mahidol University. Those collections are permitted and legal. Voucher specimens of *C. comosa* and *C. latifolia* were deposited at Sireeruckhachat Nature Learning Park and assigned as PBM005645 and PBM005639, respectively. All the experiments were performed in accordance with relevant guidelines and regulations.

Nineteen crude drug samples claimed to be *C. comosa* and *C. latifolia* were purchased from various herbal markets. They were purchased in different forms, as dried herbs and powders. Sample ID of 19 crude drugs in our collection was provided in Table S8.

Chloroplast genome sequencing. Intact chloroplasts were isolated from 10 g of fresh leaves of *C. comosa* and *C. latifolia* using 40% (*v/v*) Percoll solution and a Chloroplast Isolation Kit according to the manufacturer's protocol (Sigma–Aldrich, USA). Subsequently, genomic DNA (gDNA) was extracted from the chloroplasts using a DNeasy Plant Mini Kit (QIAGEN, Germany). The quality of the chloroplast genomic DNA was verified by measurement of the A_{260}/A_{280} ratio by a NanoDrop One UV–Vis Spectrophotometer (Thermo Scientific, USA) and precisely quantified via a Qubit 3 Fluorometer (Invitrogen, USA). Whole cp genome shotgun libraries were prepared using the SparQ Frag & DNA Library Prep Kit (Quantabio, USA). The average fragment length of the constructed DNA libraries was measured using an Agilent 2100 Bioanalyzer (GMI, USA). Paired-end sequencing was carried out on an Illumina NextSeq 500 (Illumina, USA).

Chloroplast genome assembly and annotation. The raw Illumina paired-end whole cp genome data of *C. comosa* and *C. latifolia* were trimmed and assembled into contigs using FASTP[7] and GetOrganelle[15], respectively. All paired-end reads were mapped to the assembled cp genomes with over 28 ⊠ coverage. The automatic annotator GeSeq was used to annotate the cp genome with BLAST searches against the cp genomes of sibling species[42]. The online program OGDRAW[12] was used to draw the circular cp genome map and then manually edited. The final assembly was submitted to GenBank (https://www.ncbi.nlm.nih.gov) with corresponding accession numbers.

Codon usage analysis. The cp genomes of *C. comosa* and *C. latifolia* were analyzed for relative synonymous codon usage (RSCU) in protein-coding genes using Mega-X software[20].

Repeat element analysis. The cp genome sequences (Table S1) belonging to plants of five different genera in the family Zingiberaceae, *Curcuma*, *Zingiber*, *Hedychium*, *Kaempferia*, and *Amomum*, were retrieved from the NCBI database for repeat element analysis along with our authentic *C. comosa* and *C. latifolia* cp sequences. The MIcroSAtellite (MISA)[1] program was used for simple sequence repeat (SSR) analysis with the minimum number of repeats set to 10, 5, 4, 3, 3 and 3 for mono-, di-, tri-, tetra-, penta- and hexa-nucleotides, respectively. REPuter software[21] was used to detect the size and location of long repeat sequences, including forward, reverse, complement and palindromic repeat units. Parameters were set as a Hamming distance of 3, maximum computed repeats of 150 and minimal repeat size of 30.

Sequence divergence analysis. The cp genomes (Table S1) were compared and analyzed for variable regions by using the mVISTA tool in Shuffle-LAGAN mode[11]. The annotated cp genome of *C. comosa* was used as the reference. Boundaries between the LSC, SSC and IR regions were manually defined to detect the variation in gene rearrangement between regions. The cp genomes (Table S5) belonging to plants of nine different genera in the family Zingiberaceae, *Curcuma*, *Zingiber*, *Hedychium*, *Kaempferia, Amomum*, *Alpinia, Wurfbainia, Stahlianthus* and *Lanxangia*, were retrieved from the NCBI database for sequence divergence analysis. DnaSP v.6[35] was used to calculate nucleotide variability (Pi). The parameters for sliding window analysis were a 600 bp window length and 200 bp step size.

SNP and indel detection. The cp genomes (Table S1) were aligned using MAFFT v.7.0.[17], and the sequences were edited using Mega-X software[20]. The SNP/indel was detected by using DnaSP v.6[35]. The cp genome of *C. comosa* was used as the reference.

Species-specific DNA markers for discrimination of C. comosa **from** C. latifolia **and other related species.** Three variable sequence regions, *ndh*A, trnT-*trn*L, and *ndh*C-*trn*V, from the twenty *Curcuma* spp. were aligned using MAFFT v.7.0.[17] (Table S5). The specific primers were designed based on the conserved sequence encompassing the variable sites of *C. comosa* and *C. latifolia* (Figure S1).

Testing the authenticity of crude drugs. The gDNA of nineteen samples (Table S8) was amplified by PCR with our specific primers (Table S9). PCR was performed in a volume of 12.5 µL containing 20-50 ng template gDNA, 1 ☒ PCR buffer, 3 mM $MgCl_2$, 0.2 mM dNTP mix, 0.5 µM of each primer and 0.5 U Platinum *Taq* DNA polymerase (Invitrogen, USA). The cycling conditions consisted of predenaturation at 94°C for 3 min followed by 30 cycles of denaturation at 94°C for 30 sec, annealing at 55°C (*trn*T-*trn*L and *ndh*A) or 60°C (*ndh*C-*trn*V) for 30 sec and extension at 72°C for 20 sec (*trn*T-*trn*L) or 25 sec (*ndh*A and *ndh*C-*trn*V) followed by a final extension at 72°C for 5 min. The PCR products were purified and sequenced by Macrogen, South Korea. Sequences from amplicons were BLASTed against our authentic *ndh*A, *trn*T-*trn*L, and *ndh*C-*trn*V sequences from *C. comosa* and *C. latifolia* in MAFFT v7.0.[17].

Phylogenetic analysis. The cp genomes (Table S10) belonging to plants of six different families in Zingiberales, Musaceae, Heliconiaceae, Strelitziaceae, Cannaceae, Costaceae and Zingiberaceae, were used for phylogenetic analysis. *Typha latifolia* L., an angiosperm, was used as an outgroup. The cp genome sequences were aligned using MAFFT v7.0.[17]. Subsequently, a neighbor-joining (NJ) tree was constructed, and the relative support for the branches of the NJ tree was assessed via 1,000 bootstrap replicates.

Declarations

Acknowledgments

This research project is supported by the Second Century Fund (C2F), Chulalongkorn University. The authors would like to thank the Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok, Thailand for kindly providing the equipment.

Author contributions

S.S. (Suchada Sukrong) conceived and designed the experiments. B.B. and S.S. (Sunisa Sangvirotjanapat) was responsible for field collection and plant species authentication. B.W. performed the experiments, data analyses, and drafting the manuscript. P.P. conducted the cp genome assembly. P.P. and N.P. constructed the phylogenetic tree. B.L.H.K. and P.C.S. supported the techniques for the cp genome analysis. B.W. and S.S. (Suchada Sukrong) reviewed and edited the manuscript. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinform.* **33**, 2583–2585. https://doi.org/10.1093/bioinformatics/btx198 (2017).
2. Boonsrangsom, T. Genetic diversity of 'Wan Chak Motluk' (*Curcuma comosa* Roxb.) in Thailand using morphological characteristics and random amplification of polymorphic DNA (RAPD) markers. *S. Afr. J. Bot.* **130**, 224–230. https://doi.org/10.1016/j.sajb.2020.01.005 (2020).
3. Burapan. S., Kim. M., Paisooksantivatana. Y., Eser. B. E. & Han. J. Thai *Curcuma* species: antioxidant and bioactive compounds. *Foods* **9**, 1219. https://doi.org/10.3390/foods9091219 (2020).
4. Campbell, W. H. & Gowri, G.Codon usage in higher plants, green algae, and cyanobacteria.*Plant Physiol.* **92**, 1–11. https://doi.org/10.1104/pp.92.1.1 (1990).

5. Chen, J., Xia, N., Zhao, J., Chen, J. & Henny, R. Chromosome numbers and ploidy levels of Chinese *Curcuma* species. *Hortic. Sci.* **48**, 525–530. https://doi.org/10.21273/HORTSCI.48.5.525 (2013).
6. Chen, J., Zhao, J., Erickson, D. L., Xia, N. & Kress, W. J. Testing DNA barcodes in closely related species of *Curcuma* (Zingiberaceae) from Myanmar and China. *Mol. Ecol. Resour.* **15**, 337–348. https://doi.org/10.1111/1755-0998.12319 (2015).
7. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *J. Bioinform.* **34**, 884–890. https://doi.org/10.1093/bioinformatics/bty560 (2018).
8. Cui, Y. *et al*. Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible *Amomum* species. *Int. J. Mol. Sci.* **20**, 4040. https://doi.org/10.3390/ijms20164040 (2019).
9. Dechbumroong, P., Aumnouypol, S., Denduangboripant, J. & Sukrong, S. DNA barcoding of *Aristolochia* plants and development of species-specific multiplex PCR to aid HPTLC in ascertainment of *Aristolochia* herbal materials. *PLoS ONE* **13**, e0202625. https://doi.org/10.1371/journal.pone.0202625 (2018).
10. Fazekas, A. J. *et al*. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE* **3**, e2802. https://doi.org/10.1371/journal.pone.0002802 (2008).
11. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, 273–279. https://doi.org/10.1093/nar/gkh458 (2004).
12. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, 59–64. https://doi.org/10.1093/nar/gkz238 (2019).
13. Gui, L. *et al*. Analysis of complete chloroplast genomes of *Curcuma* and the contribution to phylogeny and adaptive evolution. *Gene* **723**, 144355. https://doi.org/10.1016/j.gene.2020.144355 (2020).
14. Hu, Z. Study on DNA barcoding and chloroplast genome of medicinal plants in Compositae. Doctoral dissertation, Doctoral thesis, Hubei University of Chinese Medicine (2012).
15. Jin, J. J. *et al*. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241. https://doi.org/10.1186/s13059-020-02154-5 (2020).
16. Joseph, R., Joseph, T. & Joseph, J. Karyomorphological studies in genus *Curcuma Linn*. *Cytologia.* **64**, 313–317. https://doi.org/10.1508/cytologia.64.313 (1999).
17. Katoh, K. & Standley, D. M. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. https://doi.org/10.1093/molbev/mst010 (2013).
18. Keeratinijakal, V. & Kongkiatpaiboon, S. Distribution of phytoestrogenic diarylheptanoids and sesquiterpenoids components in *Curcuma comosa* rhizomes and its related species. *Rev. Bras. Farmacogn.* **27**, 290–296. https://doi.org/10.1016/j.bjp.2016.12.003 (2017).
19. Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. Use of DNA barcodes to identify flowering plants. *PNAS.* **102**, 8369–8374. https://doi.org/10.1073/pnas.0503123102 (2005).
20. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
21. Kurtz, S. *et al*. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29**, 4633–4642. https://doi.org/10.1093/nar/29.22.4633 (2001).
22. Li, B., Lin, F., Huang, P., Guo, W. & Zheng, Y. Complete chloroplast genome sequence of *Decaisnea insignis*: Genome organization, genomic resources and comparative analysis. *Sci. Rep.* **7**, 10073. https://doi.org/10.1038/s41598-017-10409-8 (2017).
23. Li, D. M., Zhao, C. Y. & Liu, X. F. Complete chloroplast genome sequences of *Kaempferia galanga* and *Kaempferia elegans*: molecular structures and comparative analysis. *Molecules* **24**, 474. https://doi.org/10.3390/molecules24030474 (2019).
24. Liu, Q., Dou, S., Ji, Z. & Xue, Q. Synonymous codon usage and gene function are strongly related in *Oryza sativa*. *Biosystems* **80**, 123–131. https://doi.org/10.1016/j.biosystems.2004.10.008 (2005).
25. Liang, H. & Chen, J. Comparison and phylogenetic analyses of nine complete chloroplast genomes of Zingibereae. *Forests* **12**, 710. https://doi.org/10.3390/f12060710 (2021).

26. Liang, H. *et al*. The complete chloroplast genome sequences of 14 *Curcuma* species: insights into genome evolution and phylogenetic relationships within Zingiberales. *Front. Genet.* **11**, 802. https://doi.org/10.3389/fgene.2020.00802 (2020).
27. Maknoi, J. Taxonomy and phylogeny of the genus *Curcuma* L. (Zingiberaceae) with particular reference to its occurrence in Thailand. Doctoral dissertation, Doctoral thesis, Prince of Songkla University (2006).
28. Mascarello, M. *et al*. Genome skimming reveals novel plastid markers for the molecular identification of illegally logged *African timber* species. *PLoS ONE* **16**, e0251655. https://doi.org/10.1371/journal. pone.0251655 (2021).
29. Minami, M. *et al*. Identification of *Curcuma* plants and curcumin content level by DNA polymorphisms in the *trn*S-*trnf*M intergenic spacer in chloroplast DNA. *J. Nat. Med.* **63**, 75–79. https://doi.org/10.1007/s11418-008-0283-7 (2009).
30. Paisooksativatana, Y. & Thepsen, O. Phenetic relationships of some Thai *Curcuma* species (Zingiberaceae) based on morphological, palynological and cytological evidence. *Thai J. Agric. Sci.* **34**, 47–57 (2001).
31. Park, I., Song, J. H., Yang, S. & Moon, B. C. Comparative analysis of *Actaea* chloroplast genomes and molecular marker development for the identification of authentic *Cimicifugae* rhizoma. *Plants* **9**, 157. https://doi.org/10.3390/plants9020157 (2020).
32. Pimkaew, P., Chuncharunee, A., Suksamsarn, A. & Piyajuturawat, P. Evaluation on toxicology of *Curcuma latifolia* Rosc. *Thai J. Toxicol.* **23**, 193–196 (2008).
33. Piyachaturawat, P., Ercharuporn, S. & Suksamrarn, A. Uterotrophic effect of *Curcuma comosa* in rats. *Int. J. Pharmacogn.* **33**, 334–338. https://doi.org/10.3109/13880209509065388 (1995a).
34. Piyachaturawat, P., Ercharuporn, S. & Suksamrarn, A. Oestogenic activity of *Curcuma comosa* extract in rats. *Asia Pacific J. Pharmacol.* **10**, 121–126 (1995b).
35. Rozas, J. *et al*. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **34**, 3299–3302. https://doi.org/10.1093/molbev/msx248 (2017).
36. Sasikumar, B. Genetic resources of Curcuma: diversity, characterization and utilization. *Plant Genet. Resour.* **3**, 230–251. https://doi.org/10.1079/PGR200574 (2005).
37. Sihanat, A., Theanphong, O. & Rungsihirunrat, K. Assessment of phylogenetic relationship among twenty *Curcuma* species in Thailand using amplified fragment length polymorphism marker. *J. Adv. Pharm. Technol. Res.* **11**, 134–141 (2020).
38. Siriluck, I., Ratchanok, T., Worakij, H. & Thitamin, K. Identification of 24 Species of Zingiberaceae in Thailand Using ISSR Technique. *Thai J. Agric. Sci.* **47**, 1–6 (2014).
39. Soontornchainaksaeng, P. & Jenjittikul, T. Chromosome number variation of phytoestrogen-producing *Curcuma* (Zingiberaceae) from Thailand. *J. Nat. Med.* **64**, 370–377. https://doi.org/10.1007/s11418-010-0414-9 (2010).
40. Srivastava, D. & Shanker, A. Identification of simple sequence repeats in chloroplast genomes of Magnoliids through bioinformatics approach. *Interdiscip. Sci. Comput. Life Sci.* **8**, 327–336. https://doi.org/10.1007/s12539-015-0129-4 (2016).
41. Suksamrarn, A. *et al*. Diarylheptanoids, new phytoestrogens from the rhizomes of *Curcuma comosa*: isolation, chemical modification and estrogenic activity evaluation. *Bioorgan. Med. Chem.* **16**, 6891–6902. https://doi.org/10.1016/j.bmc.2008.05.051 (2008).
42. Tillich, M. *et al*. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, 6–11. https://doi.org/10.1093/nar/gkx391 (2017).
43. Vinitha, M. R., Kumar, U. S., Aishwarya, K., Sabu, M. & Thomas, G. Prospects for discriminating Zingiberaceae species in India using DNA barcodes. *J. Integr. Plant Biol.* **56**, 760–773. https://doi.org/10.1111/jipb.12189 (2014).
44. Weerachayaphorn, J. *et al*. A protective effect of *Curcuma comosa* Roxb. on bone loss in estrogen deficient mice. *J. Ethnopharmacol.* **137**, 956–962. https://doi.org/10.1016/j.jep.2011.06.040 (2011).
45. Winuthayanon, W. *et al*. Diarylheptanoid phytoestrogens isolated from the medicinal plant *Curcuma comosa*: biological actions in vitro and in vivo indicate estrogen receptor-dependent mechanisms. *Environ. Health Persp.* **117**, 1155–1161. https://doi.org/10.1289/ehp.0900613 (2009a).
46. Winuthayanon, W. *et al*. Estrogenic activity of diarylheptanoids from *Curcuma comosa* Roxb. requires metabolic activation. *J. Agric. Food Chem.* **59**, 840–845. https://doi.org/10.1021/jf802702c (2009b).

47. Wu, L. *et al*. Plant super-barcode: a case study on genome-based identification for closely related species of *Fritillaria*. *Chin. Med.* **16**, 52. https://doi.org/10.1186/s13020-021-00460-z (2021).

48. Yang, J. B., Tang, M., Li, H. T., Zhang, Z. R. & Li, D. Z. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol. Biol.* **13**, 84. https://doi.org/10.1186/1471-2148-13-84 (2013).

49. Yik, M. H.-Y. *et al*. Differentiation of *Hedyotis diffusa* and common adulterants based on chloroplast genome sequencing and DNA barcoding markers. *Plants* **10**, 161. https://doi.org/10.3390 /plants10010161 (2021).

50. Zhitao, N. *et al*. Comparative analysis of *Dendrobium* plastomes and utility of plastomic mutational hotspots. *Sci. Rep.* **7**, 2073. https://doi.org/10.1038/s41598-017-02252-8 (2017).

51. Zhou, T. *et al*. Comparative chloroplast genome analyses of species in *Gentiana* section *Cruciata* (Gentianaceae) and the development of authentication markers. *Int. J. Mol. Sci.* **19**, 1962. https://doi.org/10.3390/ijms19071962 (2018).

Tables

Table 1. Summary of the cp genomes of *C. comosa* and *C. latifolia* and 8 related species in Zingiberaceae. LSC: large single copy, SSC: small single copy, IR: inverted repeat, PCG: protein-coding genes, tRNA: transfer RNA, rRNA: ribosomal RNA.

Species	Genome size		LSC		SSC		IR		PCGs	tRNAs	rRNAs	Total genes
	Length (bp)	%GC	Length (bp)	%GC	Length (bp)	%GC	Length (bp)	%GC				
C. comosa	162,272	36.2	87,074	34.0	15,698	29.7	29,750	41.2	87	38	8	133
C. latifolia	162,289	36.2	87,089	34.0	15,700	29.7	29,750	41.2	87	38	8	133
C. aromatica	162,243	36.2	87,109	34.0	15,634	29.8	29,750	41.2	87	38	8	133
C. longa	162,220	36.2	87,041	34.0	15,681	29.6	29,749	41.1	87	38	8	133
C. elata	162,171	36.2	87,037	34.0	15,634	29.8	29,750	41.2	87	38	8	133
C. xanthorrhiza	162,192	36.2	86,994	34.0	15,700	29.7	29,749	41.2	87	38	8	133
Zingiber officinale	162,621	36.1	87,486	33.8	15,577	29.7	29,779	41.1	87	38	8	133
Hedychium coronarium	163,949	36.1	88,581	33.8	15,808	29.5	29,780	41.1	87	38	8	133
Kaempferia galanga	163,811	36.1	88,405	33.8	15,812	29.5	29,797	41.1	87	38	8	133
Amomum krervanh	162,766	36.2	87,728	33.9	15,754	29.6	29,642	41.2	87	38	8	133

Table 2. Gene categorization and functional classification of *C. comosa* and *C. latifolia* cp genomes. Intron-containing genes are labeled with an asterisk. (x2) indicates duplicated genes in IR regions.

Gene category	Groups of genes	Name of genes
Self-replication	Ribosomal RNAs	_rrn_4.5(x2), _rrn_5(x2), _rrn_16(x2), _rrn_23(x2)
	Transfer RNAs	*_trn_A-UGC(x2), _trn_C-GCA, _trn_D-GUC, _trn_E-UUC, _trn_F-GAA, _trn_G-GCC, *_trn_G-UCC, _trn_H-GUG(x2), _trn_I-CAU(x2), *_trn_I-GAU(x2), *_trn_K-UUU, _trn_L-CAA(x2), *_trn_L-UAA, _trn_L-UAG, _trn_M-CAU, _trn_N-GUU(x2), _trn_P-UGG, _trn_Q-UUG, _trn_R-ACG(x2), _trn_R-UCU, _trn_S-GGA, _trn_S-GCU, _trn_S-UGA, _trn_T-GGU, _trn_T-UGU, _trn_V-GAC(x2), *_trn_V-UAC, _trn_W-CCA, _trn_Y-GUA, _trn_fM-CAU
	Small ribosomal subunit	_rps_2, _rps_3, _rps_4, _rps_7(x2), _rps_8, _rps_11, *_rps_12(x2), _rps_14, _rps_15, *_rps_16, _rps_18, _rps_19(x2)
	Large ribosomal subunit	*_rpl_2(x2), _rpl_14, *_rpl_16, _rpl_20, _rpl_22, _rpl_23(x2), _rpl_32, _rpl_33, _rpl_36
	DNA-dependent RNA polymerase	_rpo_A, _rpo_B, *_rpo_C1, _rpo_C2
Photosynthesis	Photosystem I	_psa_A, _psa_B, _psa_C, _psa_I, _psa_J
	Photosystem II	_psb_A, _psb_B, _psb_C, _psb_D, _psb_E, _psb_F, _psb_I, _psb_J, _psb_K, _psb_L, _psb_M, _psb_H, _psb_T, _psb_Z
	Cytochrome b/f complex	_pet_A, *_pet_B, *_pet_D, _pet_G, _pet_L, _pet_N
	ATP synthase	_atp_A, _atp_B, _atp_E, *_atp_F, _atp_H, _atp_I
	ATP-dependent protease subunit p gene	*_clp_P
	Rubisco large subunit	_rbc_L
	NADH dehydrogenase	*_ndh_A, *_ndh_B(x2), _ndh_C, _ndh_D, _ndh_E, _ndh_F, _ndh_G, _ndh_H, _ndh_I, _ndh_J, _ndh_K
Other genes	Maturase	_mat_K
	Envelope membrane protein	_cem_A
	Acetyl-CoA-carboxylase	_acc_D
	c-type cytochrome synthesis gene	_ccs_A
	Translation initiation factor	_inf_A
Genes of unknown function	Conserved open reading frames	_ycf_1(x2), _ycf_2(x2), *_ycf_3, _ycf_4

Table 3. Molecular authentication of crude drug samples from various herbal markets by using _ndh_A, _trn_T-_trn_L, and _ndh_C-_trn_V regions in the cp genome.

Sample code	Claimed species	% Identity with *ndh*A		% Identity with *trn*T-*trn*L		% Identity with *ndh*C-*trn*V		Identified species
		C. comosa	*C. latifolia*	*C. comosa*	*C. latifolia*	*C. comosa*	*C. latifolia*	
CD-01	Wan Chak Motluk (*C. comosa)*	100.00	99.52	100.00	97.18	100.00	99.32	*C. comosa*
CD-02	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-03	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-04	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-05	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-06	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-07	Wan Chak Motluk (*C. comosa)*	100.00	99.52	100.00	97.18	100.00	99.32	*C. comosa*
CD-08	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-09	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-10	Wan Chak Motluk (*C. comosa)*	98.07	98.56	95.30	98.07	93.90	92.59	not *C. comosa* or *C. latifolia*
CD-11	Wan Chak Motluk (*C. latifolia)*	93.75	94.71	95.30	98.07	95.93	95.29	not *C. comosa* or *C. latifolia*
CD-12	Wan Chak Motluk (*C. latifolia)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-13	Wan Chak Motluk (*C. latifolia)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-14	Wan Chak Motluk (*C. latifolia)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-15	Wan Chak Motluk (*C. latifolia)*	87.98	88.46	95.30	98.07	95.93	95.29	not *C. comosa*

								or *C. latifolia*
CD-16	Wan Chak Motluk (*C. comosa)*	100.00	99.52	100.00	97.18	100.00	99.32	*C. comosa*
CD-17	Wan Chak Motluk (*C. comosa)*	99.52	100.00	97.18	100.00	99.32	100.00	*C. latifolia*
CD-18	Wan Chak Motluk (*C. comosa)*	100.00	99.52	74.18	95.17	78.64	78.11	*C. comosa*
CD-19	Wan Chak Motluk (*C. comosa)*	100.00	99.52	92.49	74.88	100.00	99.32	*C. comosa*

Figures

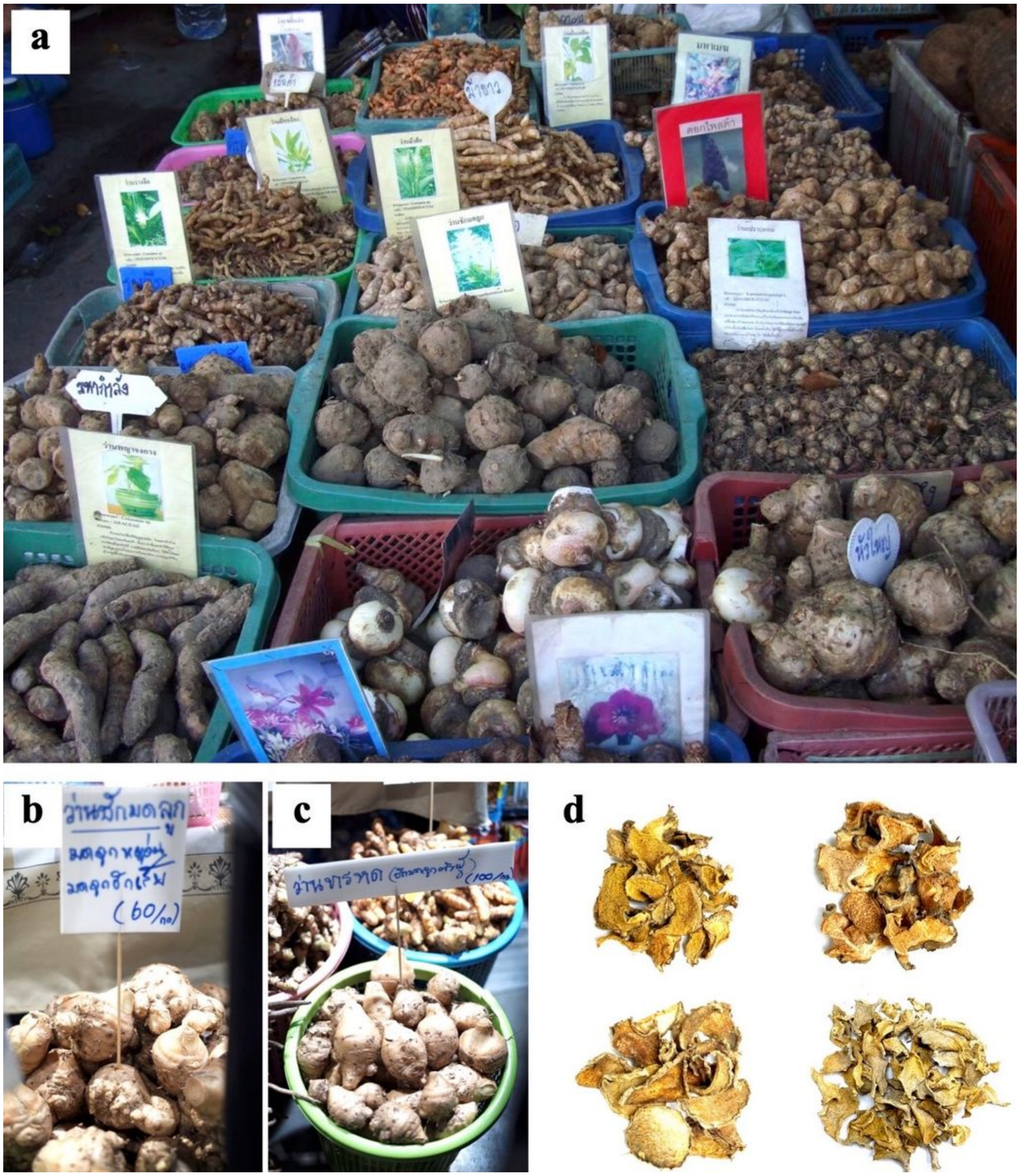

Figure 1

Various rhizomatous herbs sold at a herbal market. **(a)** Various shapes of rhizomes. **(b)** and **(c)** Fresh rhizomes labeled as “Wan Chak Motluk” and claimed to be *C. comosa* and *C. latifolia*, respectively. **(d)** Crude drugs claimed to be “Wan Chak Motluk”.

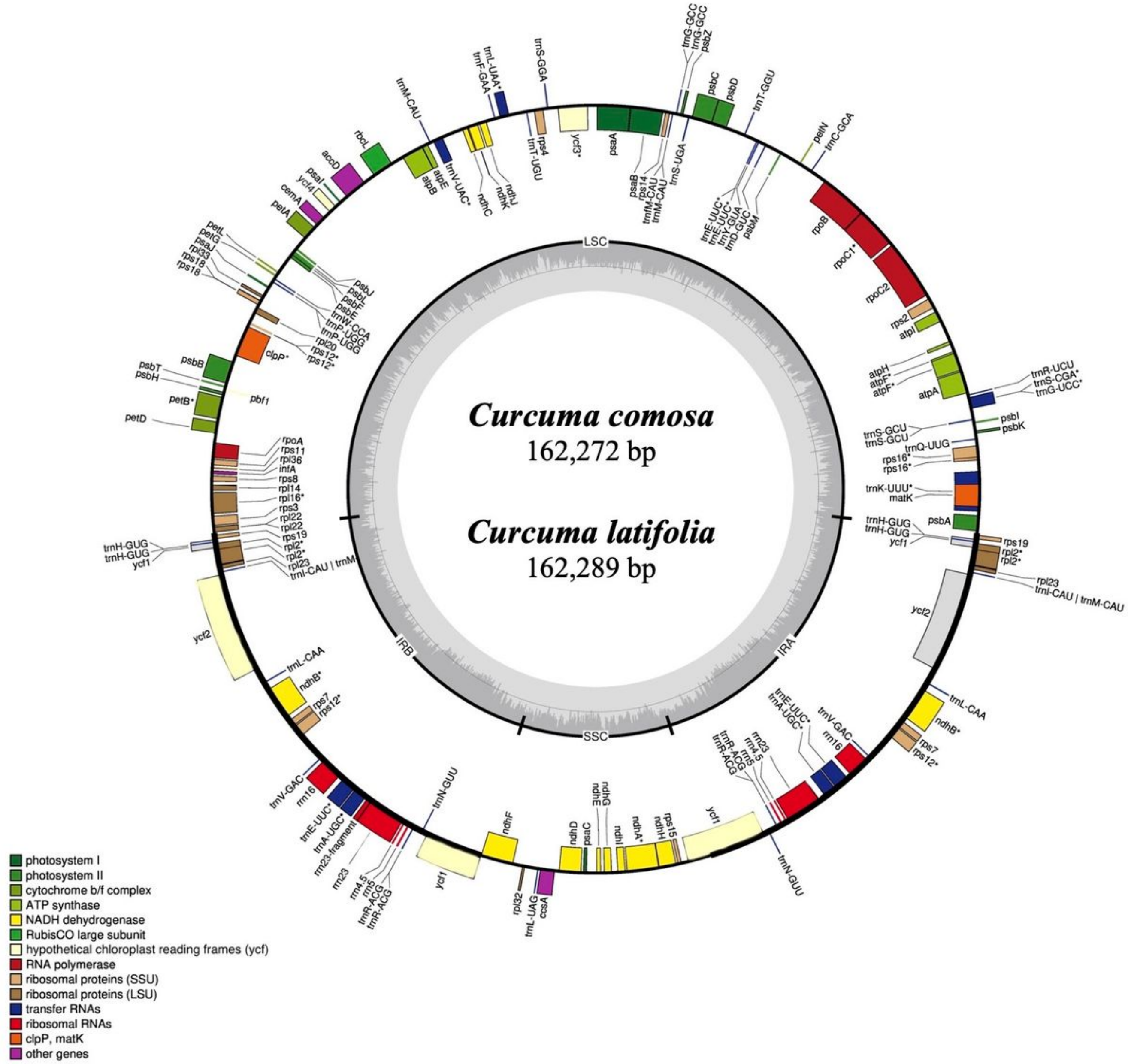

Figure 2

Gene maps of *C. comosa* and *C. latifolia* cp genomes. Genes drawn inside and outside of the circles are transcribed clockwise and counterclockwise, respectively. Genes are color coded by functional group. Asterisks indicate intron-containing genes. The darker gray and lighter gray areas in the inner circle represent the GC and AT contents, respectively.

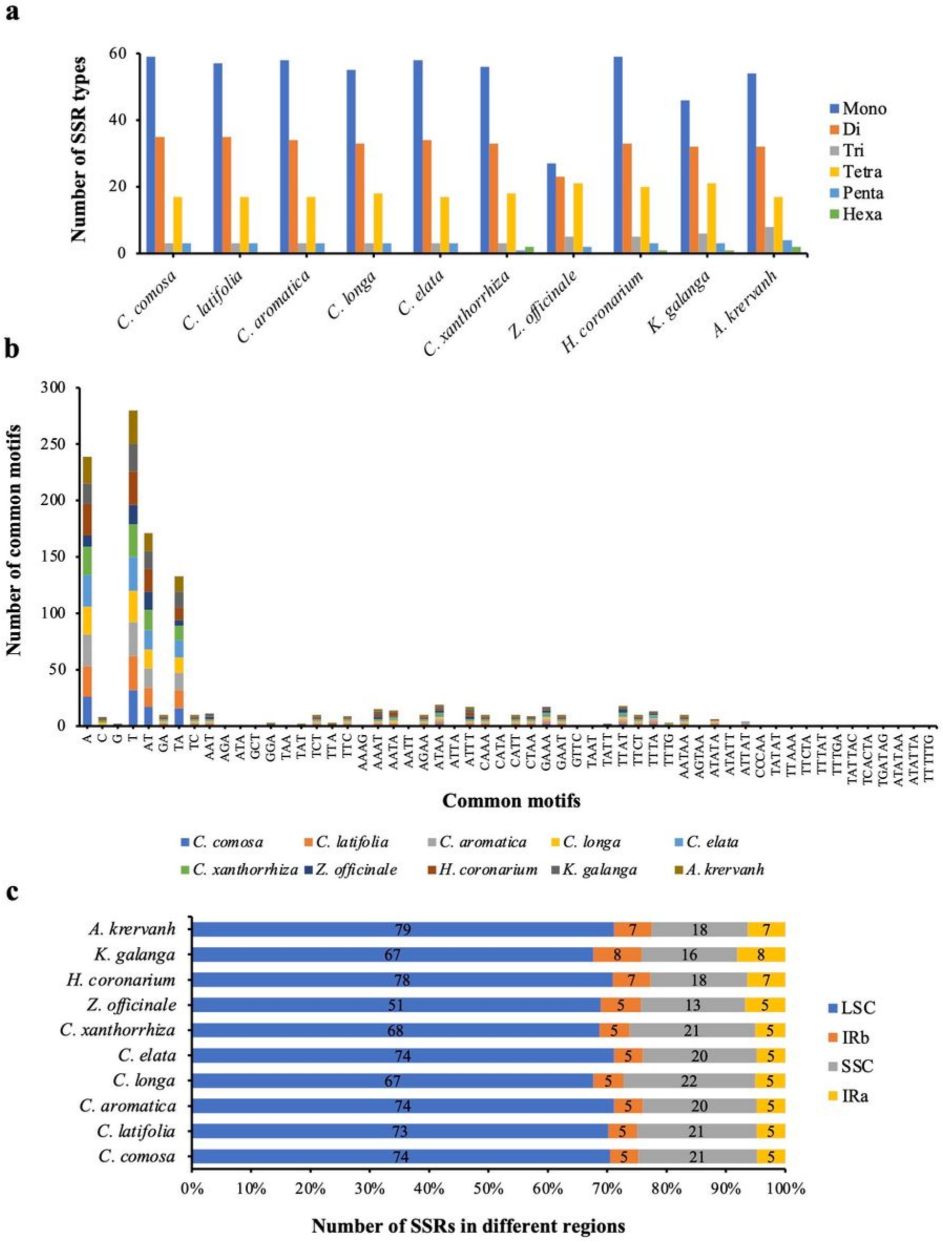

Figure 3

Simple sequence repeat (SSR) analysis of the 10 Zingiberaceae cp genomes. **(a)** Number of different SSR types. **(b)** Number of common motifs. **(c)** Number of SSRs in the LSC, SSC, and IR regions.

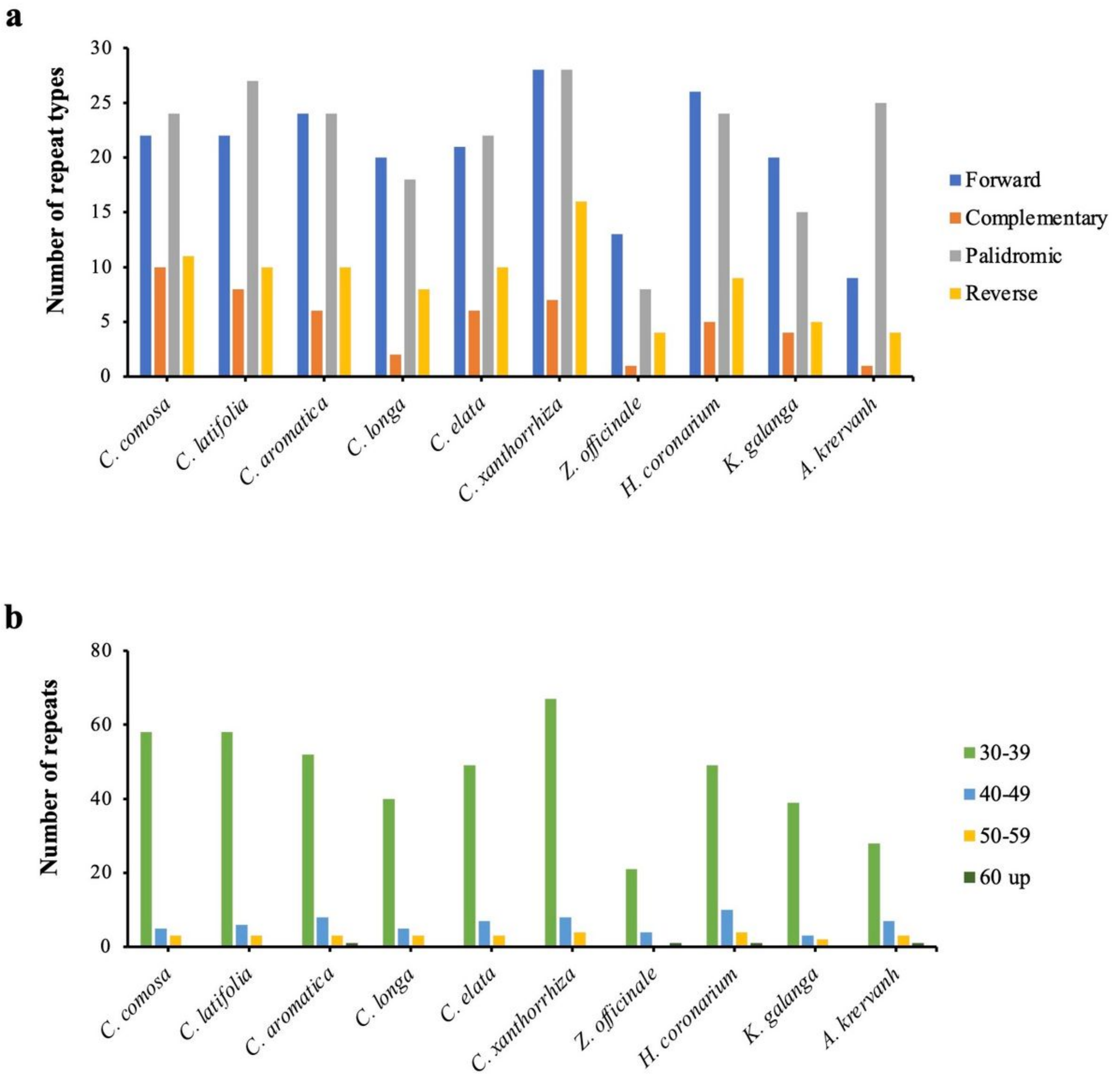

Figure 4

Long repeat sequence analysis of the 10 Zingiberaceae cp genomes. (a) Number of repeat types. (b) Number of repeats.

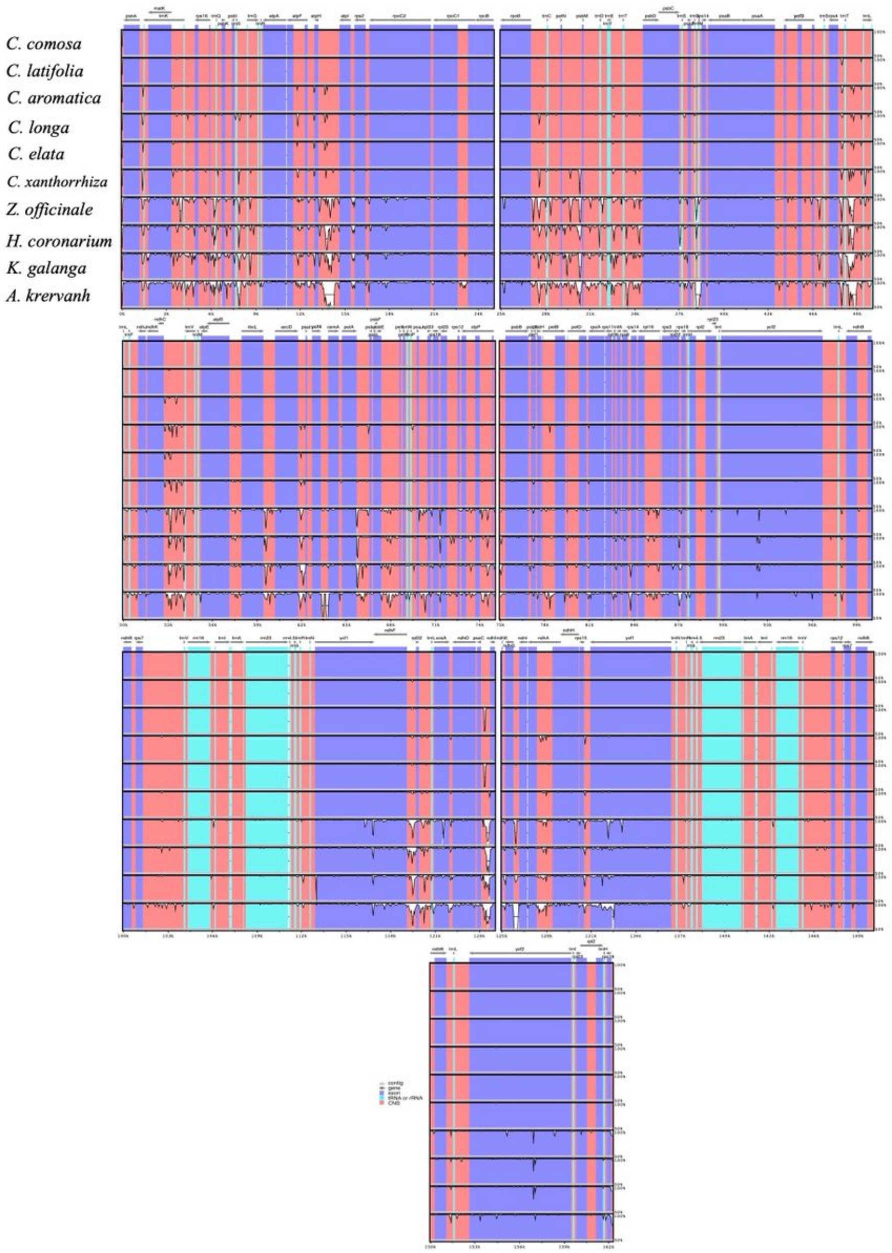

Figure 5

Sequence alignment of 10 Zingiberaceae cp genomes using mVISTA. The *C. comosa* cp genome was used as a reference. Gray arrows and thick black lines above the alignment indicate the gene orientation. Purple bars represent exons, sky-blue bars represent transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), and red bars represent noncoding sequences (NCSs). The horizontal axis indicates the coordinates within the cp genome. The vertical scale represents the percent identity, ranging from 50 to 100%. White peaks represent regions with sequence variation among the 10 species.

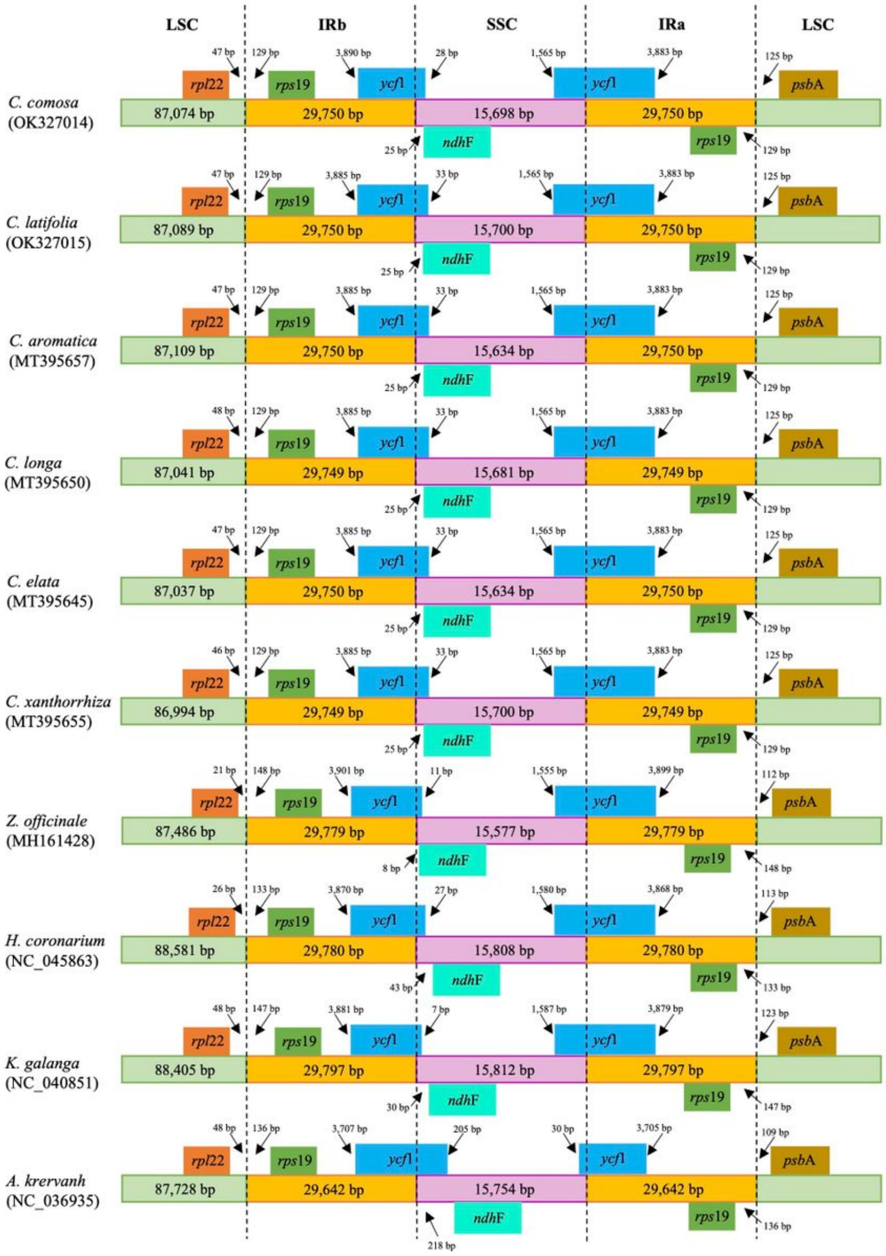

Figure 6

Comparison of LSC, SSC and IR regional boundaries among the 10 Zingiberaceae species. Numbers above the genes denote the distance between the end of the gene and the border sites. The figure is not to scale.

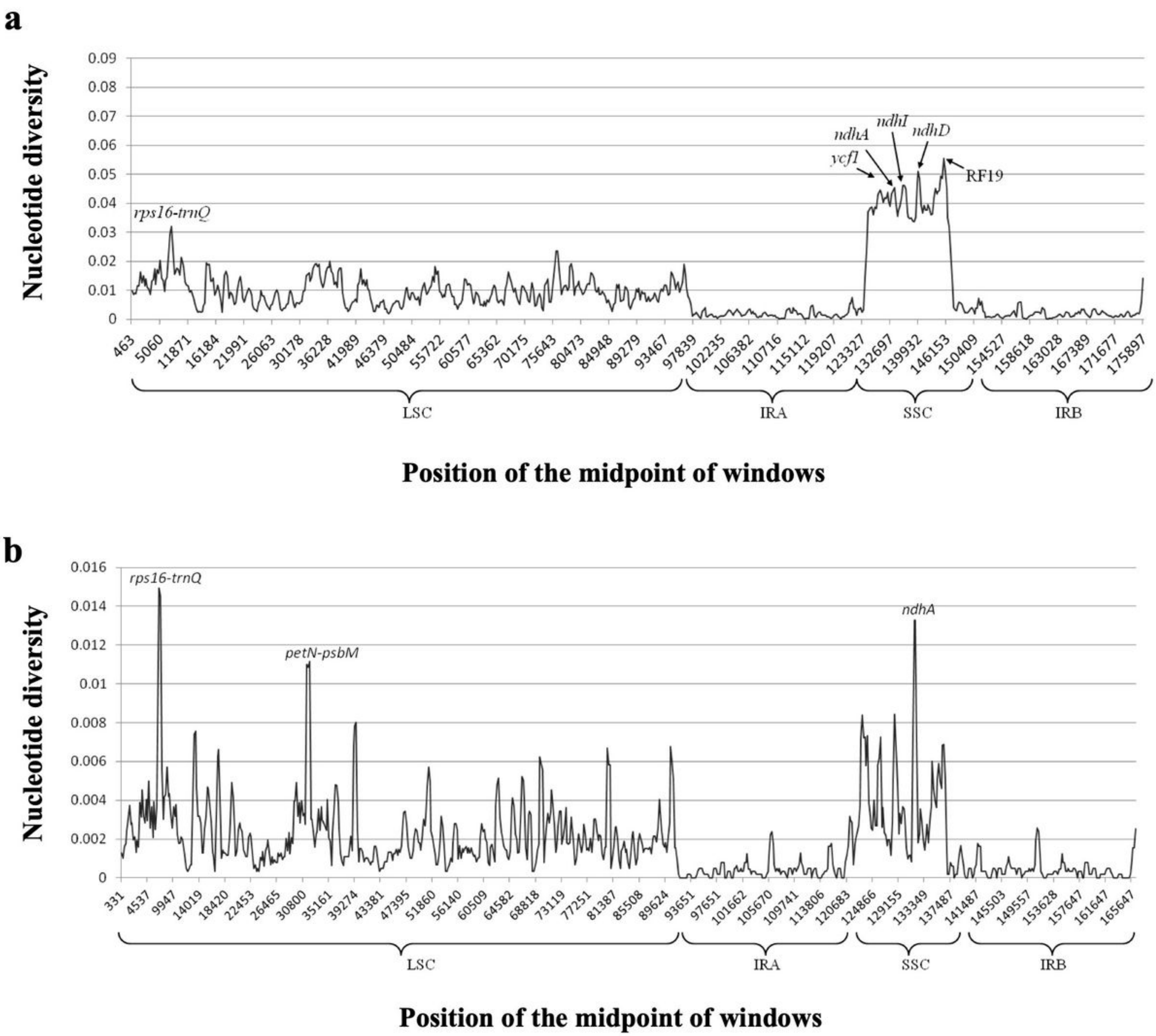

Figure 7

Sliding window analysis. **(a)** Pi among 33 plant species in Zingiberaceae. **(b)** Pi among 20 *Curcuma* species. *X*-axis: position of the midpoint of windows, *Y*-axis: nucleotide diversity of each window (Pi).

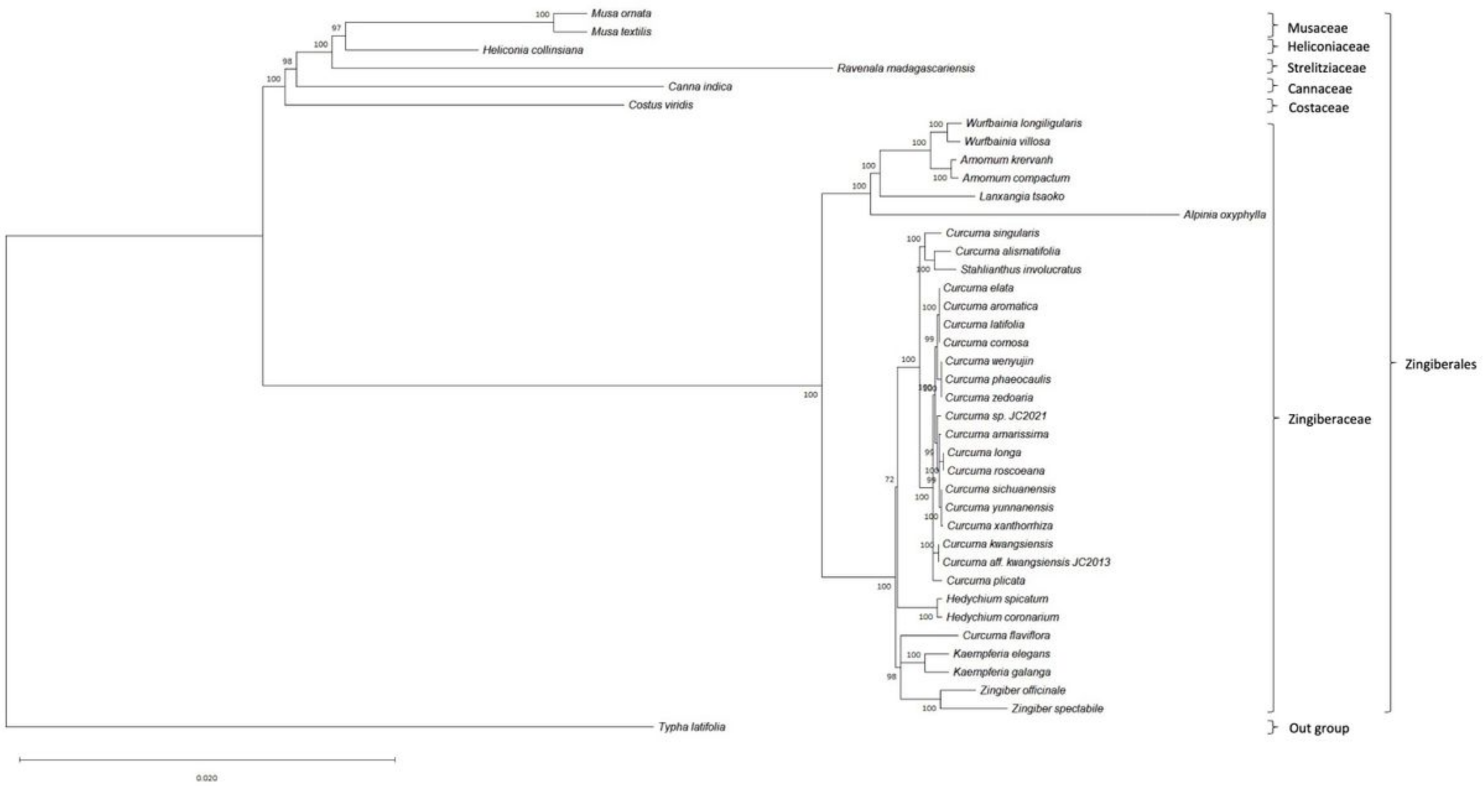

Figure 8

Phylogenetic relationships of *C. comosa* and *C. latifolia* and other related species in Zingiberales based on neighbor-joining (NJ) analysis of 40 cp genome sequences. The bootstrap values are based on 1,000 replicates. *Typha latifolia* was used as an outgroup.

Supplementary Files

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

- Supplementdataset.zip