Molecular evolution of chloroplast genomes in subfamily Zingiberoideae (Zingiberaceae)

Dong‐Mei Li*, Jie Li, Dai‐Rong Wang, Ye‐Chun Xu and Gen‐Fa Zhu*

Abstract

Background: Zingiberoideae is a large and diverse subfamily of the family Zingiberaceae. Four genera in subfamily Zingiberoideae each possess 50 or more species, including Globba (100), Hedychium (> 80), Kaempferia (50) and Zingiber (150). Despite the agricultural, medicinal and horticultural importance of these species, genomic resources and suitable molecular markers for them are currently sparse.

Results: Here, we have sequenced, assembled and analyzed ten complete chloroplast genomes from nine species of subfamily Zingiberoideae: Globba lancangensis, Globba marantina, Globba multiflora, Globba schomburgkii var. angustata, Hedychium coccineum, Hedychium neocarneum, Kaempferia rotunda 'Red Leaf', Kaempferia rotunda 'Silver Diamonds' and Zingiber recurvatum. These ten chloroplast genomes (size range 162,630–163,968 bp) possess typical quadripartite structures that consist of a large single copy (LSC, 87,172–88,632 bp), a small single copy (SSC, 15,393–15,917 bp) and a pair of inverted repeats (IRs, 29,673–29,833 bp). The genomes contain 111–113 different genes, including 79 protein coding genes, 28–30 tRNAs and 4 rRNA genes. The dynamics of the genome structures, gene contents, amino acid frequencies, codon usage patterns, RNA editing sites, simple sequence repeats and long repeats exhibit similarities, with slight differences observed among the ten genomes. Further comparative analysis of seventeen related Zingiberoideae species, 12 divergent hotspots are identified. Positive selection is observed in 14 protein coding genes, including accD, ccsA, ndhA, ndhB, psbJ, rbcL, rpl20, rpoC1, rpoC2, rps12, rps18, ycf1, ycf2 and ycf4. Phylogenetic analyses, based on the complete chloroplast‐derived single‐nucleotide polymorphism data, strongly support that Globba, Hedychium, and Curcuma I + “the Kaempferia clade” consisting of Curcuma II, Kaempferia and Zingiber, form a nested evolutionary relationship in subfamily Zingiberoideae.

Conclusions: Our study provides detailed information on ten complete Zingiberoideae chloroplast genomes, representing a valuable resource for future studies that seek to understand the molecular evolutionary dynamics in family Zingiberaceae. The identified divergent hotspots can be used for development of molecular markers for phylogenetic inference and species identification among closely related species within four genera of Globba, Hedychium, Kaempferia and Zingiber in subfamily Zingiberoideae.

Keywords: Zingiberaceae, Zingiberoideae, Chloroplast genome, Phylogeny, Divergent hotspots, Genome evolution

Background

Zingiberaceae is a family of over 1200 species that span 53 genera [1–5]. These species are found throughout the tropical and subtropical world, with their primary populations and species diversity centered in Southern and Southeast Asia [1, 3–5]. The family Zingiberaceae consists of the four recognized subfamilies of Alpinioideae, Siphonochiloideae, Tamijioideae and Zingiberoideae,
with Globba, Hedychium, Kaempferia and Zingiber belonging to Zingiberoideae [2]. These four genera each possess 50 or more species [1, 3–5]. Some of these species in subfamily Zingiberoideae have high economic value, such as G. schomburgkii and H. coronarium, ornamental and medicinal plants, K. galanga, a medicine and flavoring spice, and Z. officinale, an important crop and edible food [3–7].

It is difficult to identify Zingiberaceae plants merely based on their vegetative parts and flowers [2, 3, 5]. First, the vegetative parts of Zingiberaceae plants are often very similar, which is not suitable for species identification [3, 5]. Second, the structure of the flowers is complex, and the texture is weak, with some of them being as thin as cicada wings. Once pressed into dry specimens, it is difficult to know their original appearance, and they break and begin to decompose as soon as they are touched [5]. Past studies of Zingiberaceae have primarily concentrated on morphological classification and resources [1, 4, 5], ecology [4], cultivation and propagation [3, 4], medicinal and ornamental uses [3–7], and phylogeny [2, 8–11]. Among these phylogenetic investigations, several studies used nuclear internal transcribed spacer (ITS) and traditional chloroplast matK/trnK-matK/trnL-trnF data to explore the phylogenetic relationships within subfamily Zingiberoideae [2, 8] and within the four genera of Globba, Hedychium, Kaempferia and Zingiber [2, 9–11]. These traditional chloroplast markers have successfully identified the patterns of the evolutionary relationships within the four genera but in general, have been limited in differentiating resolutions among these four genera. Compared to traditional chloroplast markers, complete chloroplast genomes provide high resolution for relationship reconstruction within the Alpinia, Amomum, Curcuma and Zingiber genera, which allows an exploration of their phylogenetic positions in family Zingiberaceae [12–18]. Moreover, chloroplast single nucleotide polymorphism (SNP)-based phylogenetic analyses improve the phylogenetic resolution within the Alpinia, Amomum, Curcuma, Hedychium, Kaempferia, Stahlianthus and Zingiber genera in family Zingiberaceae [19–26]. However, because of the lack of complete chloroplast genomic data for the genus Globba, no studies have focused on the structural or mutational dynamics of the chloroplast genomes among the four genera of Globba, Hedychium, Kaempferia and Zingiber in subfamily Zingiberoideae.

Chloroplasts are important organelles that have their own genomes. They provide essential energy needed for plant growth and survival by converting light energy into carbohydrates through photosynthesis [27–29]. In angiosperms, chloroplast genomes typically consist of a large single copy region (LSC), a small single copy region (SSC), and two copies of inverted repeats (IRA and IRB) [27–29]. Most chloroplast genomes of flowering plants range in size from 107 kb (Cathaya argyrophylla) to 280 kb (Pelargonium) and consist of 110–130 genes, encoding ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and proteins [12–29]. In contrast with nuclear and mitochondrial genomes, chloroplast genomes are more conserved, shorter in length and more widely used in plant species identification and phylogenetic relationship analyses [12–29]. Moreover, the development of high-throughput sequencing technology has reduced the cost of sequencing and quickly accelerated comparative chloroplast genome and phylogenetic research.

In this study, we completely sequenced ten Zingiberoideae chloroplast genomes (G. lancangensis, G. marantina, G. multiflora, G. schomburgkii, G. schomburgkii var. angustata, H. coccineum, H. neocarneum, K. rotunda ‘Red Leaf’, K. rotunda ‘Silver Diamonds’ and Z. recurvatum) and compared them to eight other published chloroplast genomes from three genera, Hedychium, Kaempferia and Zingiber, within subfamily Zingiberoideae (H. coronarium, H. spicatum, K. galanga, K. elegans, Z. montanum, Z. officinale, Z. spectabile and Z. zerumbet), which were downloaded from GenBank. The primary aims of this study were as follows: (1) to compare and analyze the structure features of ten sequenced chloroplast genomes from four genera, Globba, Hedychium, Kaempferia and Zingiber; (2) to determine the sequence variation and molecular evolution among all 18 chloroplast genomes from the four genera in subfamily Zingiberoideae; and (3) to reconstruct phylogenetic relationships to verify the four genera’s relationships within subfamily Zingiberoideae.

Results
Features of ten sequenced Zingiberoideae chloroplast genomes
All ten sequenced Zingiberoideae chloroplast genomes have a typical quadripartite structure containing one large single copy (LSC), one small single copy (SSC) and two inverted repeat regions (IRA and IRB) by OGDRAW [30] and CGView tool [31] (Fig. 1, Fig. S1, Table 1). The ten sequenced Zingiberoideae chloroplast genomes size ranges from 162,630 bp (K. rotunda ‘Red Leaf’) to 163,968 bp (H. coccineum) (Table 1, Table S1, Fig. S1). They display four junction regions, namely, a separate LSC region of 87,172–88,632 bp, an SSC region of 15,393–15,917 bp, and a pair of IRs (IRA and IRB) of 29,673–29,833 bp each (Fig. 1, Fig. S1, Table 1, Table S1). The size of the G. schomburgkii var. angustata chloroplast genome (163,432 bp) is the largest among the five sequenced Globba species, with 126 bp, 658 bp, 233 bp, and 107 bp longer than G. lancangensis, G. marantina,
G. multiflora and G. schomburgkii, respectively (Table 1, Table S1). The GC content of these 10 chloroplast genomes is very similar (35.73–36.18%) (Table 1, Table S1). Specifically, the GC content in the IR regions (41.02–41.15%) is higher than that in the SSC regions (28.83–29.60%) and LSC regions (33.35–34.02%) (Table 1, Table S1). Additionally, the GC content at the third codon position (28.49–34.05%) is lower than that at the first (41.20–44.90%) and second (34.15–38.89%) positions in the protein coding genes of these 10 chloroplast genomes (Table S1). The ten sequenced chloroplast genomes contain 140–141 predicted functional genes, which consist of 87 protein coding genes, 45–46 tRNA genes, and 8 rRNA genes (Tables 1, 2, Table S2). Among these genes, a total of 111–113 different genes are detected in our 10 sequenced chloroplast genomes, including 79 protein coding genes, 28–30 tRNA genes, and 4 rRNA genes (Tables 1, 2, Table S2). Among the different protein coding genes in our 10 sequenced chloroplast genomes, 61 genes are located in the LSC regions, 12 genes are located in the SSC regions, and 8–9 genes are duplicated in the IR regions (Table 1, Table S2). Furthermore, most of the
Table 1  Comparison of ten chloroplast genomes features among the nine Zingiberoidae species studied

| Genome feature       | G. lancangensis | G. marantina | G. multiflora | G. schomburgii | G. schomburgii var. angustata | H. coccineum | H. neocarneum | K. rotunda 'Red Leaf' | K. rotunda 'Silver Diamonds' | Z. recurvatum |
|----------------------|-----------------|--------------|---------------|----------------|-------------------------------|--------------|---------------|-----------------------|-------------------------------|---------------|
| Genome size (bp)     | 163,306         | 162,774      | 163,199       | 163,325        | 163,432                       | 163,968      | 163,903       | 162,630               | 162,875                       | 163,151       |
| LSC length (bp)      | 88,545          | 87,989       | 87,994        | 88,451         | 88,556                        | 88,632       | 88,541        | 87,172                | 87,306                        | 87,780        |
| SSC length (bp)      | 15,393          | 15,425       | 15,715        | 15,525         | 15,526                        | 15,798       | 15,824        | 15,800                | 15,917                        | 15,787        |
| IR length (bp)       | 29,684          | 29,680       | 29,745/29,742 | 29,673/29,676  | 29,675/29,678                 | 29,769       | 29,769        | 29,829/29,833         | 29,826                        | 29,792        |
| GC content (%)       |                 |              |               |                |                               |              |               |                       |                               |               |
| Total genome         | 35.73           | 35.92        | 35.85         | 35.85          | 35.83                         | 36.08        | 36.08         | 36.18                 | 36.13                         | 36.12         |
| LSC                  | 33.35           | 33.58        | 33.60         | 33.51          | 33.47                         | 33.83        | 33.85         | 34.02                 | 33.97                         | 33.91         |
| SSC                  | 29.07           | 29.40        | 28.83         | 29.17          | 29.16                         | 29.56        | 29.51         | 29.60                 | 29.44                         | 29.53         |
| IR                   | 41.02/41.03     | 41.07        | 41.09/41.10   | 41.09/41.10    | 41.15                         | 41.15        | 41.08/41.09   | 41.14                 | 41.08/41.09                    |               |
| CDS                  | 36.59           | 36.68        | 36.68         | 36.68          | 36.68                         | 37.22        | 37.22         | 36.94                 | 36.96                         | 36.95         |
| Genes (total/different) | 141/113       | 140/113      | 140/113       | 141/111        | 140/113                       | 140/113      | 140/113       | 140/113               | 140/113                        |               |
| CDS (total/different) | 87/79          | 87/79        | 87/79         | 87/79          | 87/79                         | 87/79        | 87/79         | 87/79                 | 87/79                         |               |
| tRNA (total/different) | 46/30         | 45/30        | 45/30         | 46/28          | 45/30                         | 46/30        | 46/30         | 45/30                 | 45/30                         |               |
| rRNA (total/different) | 8/4           | 8/4          | 8/4           | 8/4            | 8/4                           | 8/4          | 8/4           | 8/4                   | 8/4                           |               |
| Genes with introns   | 18              | 17           | 17            | 18             | 17                            | 17           | 17            | 17                    | 17                            |               |
| Different CDS in LSC | 61              | 61           | 61            | 61             | 61                            | 61           | 61            | 61                    | 61                            |               |
| Different CDS in SSC | 12              | 12           | 12            | 12             | 12                            | 12           | 12            | 12                    | 12                            |               |
| Different CDS in IRA | 9               | 8            | 8             | 8              | 8                             | 8            | 8             | 8                     | 8                             |               |
| Different genes in IRs | 21             | 20           | 20            | 20             | 20                            | 20           | 20            | 20                    | 20                            |               |
| GenBank accession    | MT473704        | MT473705     | MT473706      | MK262735       | MT473707                      | MT47308      | MT473709      | MT473710               | MT473711                      | MT473712      |

Note: LSC large single copy region, SSC small single copy region, IR inverted repeat, CDS protein coding genes
protein coding genes and rRNAs in our 10 sequenced Zingiberoideae chloroplast genomes are similar, but there are slight differences. For instance, the chloroplast genomes of *H. coccineum* and *H. neocarneum* have the psbZ gene, while *lhbA* gene is missing in both genomes (Table 2, Fig. S1). Additionally, for tRNAs, the chloroplast genome of *H. neocarneum* and *H. coccineum* have the psbZ gene only present in the chloroplast genomes of *H. neocarneum* and *H. coccineum*, respectively; *lhbA* gene is missing in the chloroplast genomes of *H. neocarneum* and *H. coccineum*, respectively; trnS-GCU and trnT-UGU exist two gene copies only in the chloroplast genome of *G. schomburgkii*, and only once in other 9 sequenced chloroplast genomes in this study; trnS-GGA and trnT-UGU are missing in the chloroplast genome of *G. schomburgkii*.

A total of 18 genes contain introns in the chloroplast genomes of *G. lancangensis* and *G. schomburgkii*. Sixteen genes (atpF, ndhA, ndhB, petB, petD, rpl2, rpl16, rpoC1, rps12, rps16, trnA-UGC, trnG-GCC, trnI-GAU, trnK-UUU, trnL-UAA, and trnV-UCAC) contain one intron, while clpP and ycf3 each contains two introns (Table 2, Table S3). Among the 18 intron-containing genes in the chloroplast genomes of *G. lancangensis* and *G. schomburgkii*, four genes (ndhB, rpl2, trnA-UCAC and trnI-GAU) occur in both IRs, 12 genes (atpF, clpP, petB, petD, rpl16, rpoC1, rps12, trnA-UGC, trnG-GCC, trnI-GAU, trnK-UUU, trnL-UAA, and trnV-UCAC) contain one intron, while clpP and ycf3 contain two introns each (Table S3). The locations of these 17 genes

---

**Table 2** Genes present in the ten sequenced chloroplast genomes in subfamily Zingiberoideae

| Category for genes | Group of genes | Name of genes |
|--------------------|----------------|---------------|
| Photosynthesis     | Subunits of photosystem I | psaA, psaB, psaC, psaI |
|                    | Subunits of photosystem II | psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbK, psbL, psbM, psbN, psbT, | psbZ, lhbA |
|                    | Subunits of ATP synthase | atpA, atpB, atpE, atpF, atpH, atpI |
|                    | Subunits of NADH dehydrogenase | ndhA, ndh8 (x 1)*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhK |
|                    | Subunit of rubisco | rbcL |
| Self-replication   | RNA polymerase | rpoA, rpoB, rpoC1, rpoC2 |
|                    | Large subunit of ribosomal proteins | rpl2 (x 2)*, rpl4, rpl16*, rpl20, rpl22, rpl23 (x 2), rpl32, rpl33, rpl36 |
|                    | Small subunit of ribosomal proteins | rps2, rps3, rps4, rps7 (x 2), rps8, rps11, rps12 (x 2)*, rps14, rps15, rps16*, rps18, rps19 (x 2) |
|                    | Ribosomal RNAs | rrn4.5 (x 2), rrn5 (x 2), rrn16 (x 2), rrn23 (x 2) |
|                    | Transfer RNAs | trnA-UGC (x 4)*, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnM-CAU, trnG-GCC (x 2)*, trnG-GCU, trnH-GUG (x 2), trnI-GAU (x 2)*, trnI-UUU (x 2)*, trnK-CAA (x 2), trnL-UAA (x 2)*, trnL-UGC, trnN-GUU (x 2), trnP-UGG, trnQ-UGU, trnR-ACG (x 2), trnR-UCU, trnS-GCU (x 2), trnS-GGA, trnS-UGC, trnT-UGU, trnV-GAC (x 2), trnV-UCAC (x 2)*, trnW-CCA, trnY-GUA |
| Other genes        | Subunit of acetyl-CoA-carboxylase | acdA* |
|                    | c-type cytochrome synthesis gene | ccsA |
|                    | Envelope membrane protein | cemA |
|                    | Protease | clpP** |
|                    | Translational initiation factor | infA |
|                    | Maturase | matK |
| Unknown function   | Conserved open reading frames | ycf1 (x 2), ycf2 (x 2), ycf3**, ycf4 |

Note: *①*: gene with two copies; *②*: gene with four copies; *③*: gene containing one intron; *④*: gene containing two introns; ①: psbZ gene is only present in the chloroplast genomes of *H. neocarneum* and *H. coccineum*, respectively; ②: lhbA gene is missing in the chloroplast genomes of *H. neocarneum* and *H. coccineum*, respectively; ③: trnS-GCU and trnT-UGU exist two gene copies only in the chloroplast genome of *G. schomburgkii*, and only once in other 9 sequenced chloroplast genomes in this study; ④: trnS-GGA and trnT-UGU are missing in the chloroplast genome of *G. schomburgkii*.
are the same as in the two chloroplast genomes of *G. lancangensis* and *G. schomburgkii*.

**Analyses of codon usage and predicted RNA editing sites**

A total of 79 protein coding genes in all 10 sequenced chloroplast genomes in subfamily Zingiberoideae are analyzed for codon usage frequency. They comprise 26,400 (*G. schomburgkii var. angustata*) to 27,730 (*G. schomburgkii*) codons. Of the 26,400–27,730 codons, leucine (Leu) is the most abundant amino acid, with a frequency of 10.14–10.69%, followed by isoleucine (Ile) with a frequency of 8.77–8.80%, while cysteine (Cys) is the least common, with a frequency of 1.13–1.15%.
(Fig. 2a, Table S4). This phenomenon is consistent with other Zingiberaceae plant chloroplast genomes, such as Z. officinale [15], K. galanga [18], A. oxyphylla [19] and A. pumila [19]. Because of the value of relative synonymous codon usage (RSCU) > 1.00, thirty codons show codon usage bias in the 10 sequenced chloroplast genomes’ protein coding genes (Fig. 2b, Table S5). Interestingly, out of the above 30 codons, twenty-nine are A/T-ending codons (Fig. 2b, Table S5). Conversely, C/G-ending codons have RSCU values of less than one, which indicates that they are less common in the 10 sequenced chloroplast genomes’ genes. Stop codon usage is biased toward TAA (RSCU > 1.00) (Fig. 2b, Table S5). Both methionine (Met) and tryptophan (Trp) exhibit no codon bias and have RSCU values of 1.00 (Fig. 2b, Table S5).

Furthermore, a total of 61 editing sites in 22 protein coding genes are identified in G. schomburgkii, while more numbers are found in G. lancangensis (76 sites), G. marantina (80 sites), G. multiflora (79 sites), G. schomburgkii var. angustata (78 sites), H. coccineum (79 sites), H. neocarneum (79 sites), K. rotunda ‘Red Leaf’ (81 sites), K. rotunda ‘Silver Diamonds’ (81 sites), and Z. recurvatum (80 sites) (Table S6). In the 10 identified chloroplast genomes that we sequenced, the ndhB gene has the highest number of potential editing sites (11), followed by the ndhD gene (9) (Table S6). Similar to other Zingiberaceae species, such as K. galanga [18], A. pumila [19], and Z. zerumbet [26], the ndhB gene contains the highest number of editing sites. All of these editing sites are C-to-T transitions that occur at the first or second positions of the codons. Interestingly, most RNA editing sites lead to amino acid changes for hydrophobic products, such as leucine, isoleucine, tryptophan, tyrosine, valine, methionine, and phenylalanine (Table S6). Similar RNA editing features have been identified in previous studies [18, 19, 26].

Analyses of simple sequence repeats (SSRs) and long repeats
In this study, there were 221 to 258 SSRs in each sequenced chloroplast genome that ranged from 8 to 27 bp in length (Fig. 3, Tables S7, S8). We discovered 6 types of SSRs, specifically, mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide. Among these SSRs, only the chloroplast genomes of G. marantina, G. multiflora, G. schomburgkii, G. schomburgkii var. angustata, H. neocarneum, K. rotunda ‘Silver Diamonds’ and Z. recurvatum exhibited hexanucleotide repeats (Fig. 3a, Table S7). Among each sequenced chloroplast genome, mononucleotide repeats were the most frequent, with numbers ranging from 163 to 193, which accounted for 73.75–76.73% of all SSRs, followed by dinucleotide, ranging from 27 to 34 (11.02–13.92%), tetranucleotide, ranging from 15 to 23 (5.92–10.41%), trinucleotide, ranging from 5 to 9 (2.11–3.58%), pentanucleotide, ranging from 3 to 5 (1.17–1.93%), and hexanucleotide, ranging from 0 to 2 (0–0.90%) (Fig. 3a, Table S7). The majority of the mononucleotide SSRs were A/T repeats, which accounted for 94.79–97.59% of all the mononucleotide types among the ten sequenced chloroplast genomes (Fig. 3b, Table S7). In the dinucleotide category, the AT/AT repeats were the most abundant type, with 48.14–66.67% of the loci in this category (Fig. 3b, Table S7). SSRs were more frequently located in the LSC regions (138–173 loci, 60.33–66.53%) than in the SSC regions (37–46 loci, 16.44–19.00%) and IR regions (44–50 loci, 17.39–22.22%) of the ten sequenced chloroplast genomes (Fig. 3c, Table S7). Likewise, SSRs were analyzed in the protein coding regions (exon, protein coding exon), intron regions and intergenic regions of the ten sequenced chloroplast genomes, which indicated that these ten sequenced chloroplast genomes contained 100 to 120 SSRs in intergenic regions, 19 to 25 SSRs in introns, and 44 to 60 SSRs in protein coding regions (Fig. 3d, Table S8).

Additionally, ten sequenced chloroplast genomes had 532 long repeats that consisted of 192 forward repeats, 24 complement repeats, 59 reverse repeats, and 257 palindromic repeats (Fig. 4a, Table S9). Among the ten sequenced genomes, G. lancangensis had the smallest number (48), and H. neocarneum had the largest number of long repeats (70) (Fig. 4a, Table S9). The number of forward repeats varied between 12 (G. schomburgkii var. angustata) and 28 (H. coccineum), the number of complement repeats varied from 1 (G. lancangensis and K. rotunda ‘Silver Diamonds’) to 5 (G. schomburgkii var. angustata), the number of reverse repeats varied between 3 (H. coccineum and K. rotunda ‘Silver Diamonds’) and 9 (H. neocarneum), and the number of palindromic repeats...
**Fig. 3** (See legend on previous page.)

(a) Number of SSRs by species and type.

(b) Number of SSRs by nucleotide type and species.

(c) Distribution of SSRs in different species.

(d) Distribution of SSRs in different regions of the genome.

Legend:
- G. lancangensis
- G. marantina
- G. multiflora
- H. neocarneum
- H. coccineum
- K. rotunda ‘Silver Diamonds’
- K. rotunda ‘Red Leaf’

- Mononucleotide
- Dinucleotide
- Trinucleotide
- Tetranucleotide
- Pentanucleotide
- Hexanucleotide
varied from 22 (G. schomburgkii) to 30 (H. coccineum and H. neocarneum) (Fig. 4a, Table S9). There was no complement repeat in Z. recurvatum (Fig. 4a, Table S9). Long repeats with 30–39 bp were found to be the most common in the ten sequenced chloroplast genomes (Fig. 4b, Table S9). Long repeats with a length of 30 bp were the most common (148), and those with lengths of 31 bp and 32 bp were the second (74) and third (63), most common, respectively (Table S9). Collectively, the number, length and distribution of these long repeats varied from one species to another among the nine tested species in the current study.

Contraction and expansion of inverted repeats (IRs)
A comprehensive comparison at the LSC/IRs/SSC boundaries was performed among the 17 Zingiberoideae species, including 5 Globba species, 4 Hedychiun species, 3 Kaempferia species and 5 Zingiber species (Fig. 5). Although the IR region of the 17 Zingiberoideae species’ chloroplast genomes was highly conserved, structure variation was still found in the IR/SC boundary regions. Within the 17 Zingiberoideae species, the rpl22 and rps19 genes were located in the boundaries of the LSC/IRB regions, except for Z. spectabile, in which there were trnM and ycf2 genes, and the rpl22/rps19 gene was absent in the boundaries of the LSC/IRB regions (Fig. 5). There were 20–125 bp between rpl22 and the LSC/IRB borders within the rest of the 16 Zingiberoideae species, and the distance between rps19 and the LSC/IRB boundary ranged from 123 bp to 173 bp (Fig. 5).

Ψycf1-ndhF genes were located at the boundaries of the IRB/SSC regions in all 17 Zingiberoideae species. The IRB/SSC borders of 8 species (G. lancangensis, G. schomburgkii var. angustata, H. coccineum, H. neocarneum, K. elegans, K. rotunda ‘Red Leaf’, K. rotunda ‘Silver Diamonds’, Z. recurvatum and Z. spectabile) were all situated adjacent to the end of Ψycf1, while 2 species (H. spicatum and Z. officinale) were found 533 bp and 563 bp distances between Ψycf1 and...
Fig. 5 Comparisons of LSC, SSC and IR regions boundaries among 18 chloroplast genomes in subfamily Zingiberoideae. ψ, pseudogenes. The figure was not to scale with respect to sequence length, and only showed relative changes at or near the IR/SC borders. The ten sequenced chloroplast genomes in this study were marked in bold.
and the IRB/SSC boundary, respectively. \(\Psi_{ycf1}\) expanded into the SSC regions in 7 species, namely, \(G. marantina\), \(G. multiflora\), \(G. schomburgkii\), \(H. coronarium\), \(K. galanga\), \(Z. montanum\) and \(Z. zerumbet\), for 1 bp, 1 bp, 13 bp, 27 bp, 8 bp, 8 bp, and 8 bp, respectively (Fig. 5). There were 33 bp, 37 bp, 14 bp, 43 bp, 23 bp, 31 bp and 31 bp between the \(Z. zerumbet\) and \(Z. montanum\), 14 bp, 43 bp, 23 bp, 31 bp and 31 bp between the ndhF and \(\Psi_{ycf1}\) border in \(G. marantina\), \(G. multiflora\), \(G. schomburgkii\), \(H. coronarium\), \(K. galanga\), \(Z. montanum\) and \(Z. zerumbet\), respectively (Fig. 5). In 9 species (\(G. schomburgkii\) var. \(angustata\), \(H. cocinea\), \(H. neocarneum\), \(H. spicatum\), \(K. elegans\), \(K. rotunda\) ‘Silver Diamonds’, \(Z. officinale\), \(Z. recurvatum\) and \(Z. spectabile\)), the distances between species, a 94–219 bp distance was observed between the \(psbA\) from 123 bp to 173 bp (Fig. 5). For all 17 Zingiberoideae distances between \(G. lancangensis\) and the IRA/LSC border ranged from 1 bp to 13 bp, 27 bp, \(Z. spectabile\) and \(Z. officinale\)), the distances between \(G. multiflora\), \(G. marantina\) and \(Z. spectabile\), the distances between \(G. montanum\) and \(Z. zerumbet\), respectively (Fig. 5). The SSC/IRA boundary was situated in the \(ycf1\) coding region, which crossed into the IRA region in all 17 Zingiberoideae species. However, the length of \(ycf1\) in the IRA region varied among the 17 Zingiberoideae species from 123 bp to 173 bp (Fig. 5). For all 17 Zingiberoideae species, except for \(G. lancangensis\) and \(Z. officinale\), the \(ycf1\) gene and the IRA/LSC border (Fig. 5).

Comparative genomic and nucleotide diversity analyses

Multiple alignments of 18 Zingiberoideae chloroplast genomes coming from 17 species of four genera, \(Globba\), \(Hedychiurn\), \(Kaempferia\) and \(Zinger\) were compared by using CGView and mVISTA, with the annotated \(G. lancangensis\) genome sequence as the reference (Figs. 1 and 6). The mVISTA comparison showed that the LSC and SSC regions were more divergent than the two IR regions and that a higher divergence was found in non-coding regions than in coding regions (Fig. 6). The main divergences for the coding regions were \(accD\), \(matK\), \(psaI\), \(rpl32\), \(rpl33\), \(rps15\), \(rps16\) and \(ycf1\). For the non-coding regions, strongly divergent regions were \(psbA-\text{trnK}-\text{UUGI}\), \(rps16-\text{trnQ}-\text{UUGI}\), \(\text{trnS-GCU-trnG-GCC}\), \(\text{atpl-athP}\), \(\text{psbM-trnD-GUCC}\), \(\text{accD-psaI}\), \(\text{ycf4-cemA}\), \(\text{psaI-rpl33}\), \(\text{trnD-GUCC-trnY-GUAA}\), \(\text{ndhF-rpl32}\), \(\text{rpl32-trnL-UAG}\), \(\text{cssA-ndhD}\) and \(\text{rps15-ycf1}\) (Fig. 6). The CGView result also indicated that the two IR regions were less divergent than the LSC and SSC regions, and major differences originated from LSC and SSC regions (innermost fourth color ring to outwards 21st ring in Fig. 1). Compared to the chloroplast genome of \(G. lancangensis\) (innermost fourth color ring in Fig. 1), the other 17 chloroplast genomes shared five divergent regions in LSC (\(\text{trnS-GCU-trnG-GCC}\), \(\text{atpl-athP}\), \(\text{trnT-GUCC-psbD}\), \(\text{psbM-trnD-GUCC}\), \(\text{and trnT-GUCC-trnL-UAA}\)), one region in SSC (\(\text{cssA-ndhD}\)) and one region in IRA (\(\text{ycf1}\)).

Furthermore, the nucleotide diversity (Pi) values were analyzed by DnaSP to test divergence level within different regions among the 18 Zingiberoideae chloroplast genomes from four genera, \(Globba\), \(Hedychiurn\), \(Kaempferia\) and \(Zinger\) (Fig. 7). In the protein coding regions, the Pi values for each locus ranged from 0 to 0.00624 and had an average value of 0.00496 (Table S10a). Of these protein coding regions, ten regions (\(\text{rps16-CD52, clpP-CD52, clpP-CD53, rps3, ndhF, rpl32, rpl33, ndha-CD51, ycf1-D2 and rps19-D2}\) exhibited remarkably high values (\(\text{Pi} > 0.01313\); Fig. 7a). For the intron and intergenic regions, the Pi values ranged from 0 to 0.07066 and had an average of 0.01327 (Table S10b). Among these intron and intergenic regions, ten most divergent regions of \(\text{rps16-CD51-trnQ-UUGI, psbl-trnS-GCU, psbC-trnS-UGA, petA-psbJ, petA-psbJ, trnl-CAL-ycf2, trnl-UAG-csaA, csaA-ndhD, psaC-ndhE and ndhH-rps15}\) with Pi values ranging from 0.03300 to 0.07066, were identified (Fig. 7b). Furthermore, using the region length ≥ 150 bp and Pi value ≥ 0.02018 for selection potential molecular markers, 23 regions were obtained: \(\text{accD-psaI, csaA-ndhD, ndhF-rpl32, petA-psbJ, psaC-ndhE, psaI-rpl33, psbA-trnK-UUGI-CD52, rbcL-accD, rpl32-trnL-UAG, rps18-rpl20, rps15-ycf1, rps16-CD51-trnQ-UUGI, psbM-trnD-GUCC, trnD-GUCC-trnY-GUAA, trnE-UUC-trnT-GUUI, trnG-UCC-trnF-CALI, trnM-CAL-athP, trnK-UUGI-CD51-rps16-CD52, trnS-GCU-trnG-GCC-CD51, trnT-UUGI-trnL-UAA-CD51, trnT-GUUI-psbD, trnW-CCA-trnP-UUGI}\) and \(\text{ycf4-cemA}\) (Table S10b).

Combing the results of mVISTA, CGView and DnaSP, 33 regions were extracted and constructed maximum
Fig. 6 (See legend on previous page.)
likelihood (ML) trees to identify 17 Zingiberoideae spe-
cies among four genera of *Globba*, *Hedychium*, *Kaempferia* and *Zingiber*. The resolution power of the divergent regions depended on the number of species successfully identified in ML trees. If the bootstrap value of the node was less than 50, species on the ML tree were not counted. Finally, based on the ML trees, 12 regions (*ndhF*, *ycf1*, *trnK-UUU-CDS1-rps16-CDS2*, *psaL-rpl33*, *ycf4-cemA*, *trnT-UGU-trnL- UAA-CDS1*, *trnT-GGU-psbD*, *rpl32-trnL-UAG*, *psbM-trnD-GUC*, *ndhF-rpl32*, *rps15-ycf1* and *ccsA-ndhD*) showed relatively high resolution power at genus level. *ycf1* had the highest resolution power of 100%, followed by *trnK-UUU-CDS1-rps16-CDS2* at 88.89%, *ycf4-cemA* at 88.89%, *trnT- GGU-psbD* at 88.89%, *trnT-UGU-trnL-UAA-CDS1* at 88.89%, *ndhF* at 83.33%, *psaL-rpl33* at 83.33%, and *rps15-ycf1* at 77.78%

![Fig. 7 Nucleotide diversity (Pi) values of various regions in 18 chloroplast genomes in subfamily Zingiberoideae.](image)

**a** Protein coding regions. Peak regions with a Pi value of > 0.0128 were labeled with loci tags of genic names. **b** Intron and intergenic regions. Peak regions with a Pi value of > 0.033 were labeled with loci tags of intergenic region names.
(Table 3, Fig. S2). Among these 8 regions, ycf1 was shared by four genera of Globba, Hedychium, Kaempferia and Zingiber; ndhF, trnK-UGU-CDS1-rps16-CDS2, psal-rpl33, and trnT-UGU-trnL-UAA-CDS1 were shared by Globba, Kaempferia and Zingiber; trnT-GGU-psbD was shared by Hedychium, Kaempferia and Zingiber; rps15-ycf1 was shared by Globba and Kaempferia; and ycf4-cemA was shared by Globba and Zingiber (bootstrap values ≥ 50%) (Table 3). One DNA universal barcode, matK, exhibited resolution power of 77.78%. While four regions, psbM-trnD-GUC, ndhF-rpl32, rpl32-trnL-UAG and ccsA-ndhD, had the resolution power less than the resolution power of matK: psbM-trnD-GUC had the resolution power of 72.22%, ndhF-rpl32 and rpl32-trnL-UAG had the same resolution power of 66.67%, and ccsA-ndhD had the lowest resolution power of 44.44% (Table 3). However, the three regions rpl32-trnL-UAG, psbM-trnD-GUC, and ccsA-ndhD could be used as candidate DNA barcodes for Globba species (all five genera successfully differentiated with bootstrap values ≥ 71%); and ndhF-rpl32 could be used as candidate DNA barcodes for Globba (bootstrap values ≥ 56%) and Kaempferia species (bootstrap values ≥ 87%), respectively (Table 3, Fig. S2i, j, m).

Characterization of substitution rates and positive selection analyses

The nonsynonymous (Ka) and synonymous (Ks) nucleotide substitution rates of all 79 protein coding genes were analyzed across 17 Zingiberoideae species. Overall, the Ka/Ks ratios were less than 1.00 and invalid for most pairs comparison (95.37%) (Table S11a). There were 49 protein coding genes with Ka/Ks ratios greater than 1.00 and p values less than 0.05 at nucleotide level, such as accD, psbl, rbcL, rpl20, rps7, rps8, rps15, rps16, ycf1, ycf2 and so on (Table S11b). These data sets were so sophisticated and may generate false positives. To measure truly positive selection at the protein level for further, we used a Bayes empirical bayes (BEB) approach in PAML [35] to integrate over these uncertainties. The BEB method inferred that some amino acid sites of 14 protein coding genes were truly under positive selection with posterior probability greater than 0.95 (Table 4, Table S12). Among the 14 protein coding genes, rps12 showed the highest number of positive amino acids sites (40), followed by ycf1 (34) and ycf2 (20) (Table 4). The other 11 protein coding genes, accD, ccsA, ndhA, ndhB, psbl, rbcL, rpl20, rpoC1, rpoC2, rps18, and ycf4, presented 4, 4, 5, 3, 1, 8, 2, 2, 2, 1, and 2 amino acids sites truly under positive selection, respectively (Table 4). The amino acids encoded by the sites rpl20 (118), ycf1 (1341, 1433, 1452, 1453, 1528 and 1586), and ycf2 (1343) were exclusively found in G. lancangensis among all 17 Zingiberoideae species studied here (Table S12). Additionally, some amino acids positions were highly variable, such as accD (4, 9 and 299), ccsA (180), rbcL (449), rps18 (27) and ycf1 (215, 928 and 1452), which displayed three or more amino acid changes among all 17 Zingiberoideae species studied here (Table S12).

Phylogenetic inference of subfamily Zingiberoideae

To examine the phylogenetic positions of 5 Globba species, 4 Hedychium species, 3 Kaempferia species and 5 Zingiber species, and their relationships within subfamily Zingiberoideae, Bayesian inference (BI) and ML trees were constructed based on the single nucleotide polymorphism (SNP) matrix from 59 chloroplast genomes (Fig. 8, Fig. S3). These chloroplast genomes included 16 of subfamily Alpinioideae, 40 of subfamily Zingiberoideae, and 3 species (Canna indica, Costus pulverulentus, and Costus viridis) as the outgroups. The topological structures of the BI and ML trees were consistent, and were divided into two subfamilies of Alpinioideae and Zingiberoideae with strong support (posterior probability = 1.00 for the BI tree and bootstrap value = 100% for the ML tree) (Fig. 8, Fig. S3). In the BI tree, the posterior probabilities of all nodes reached 1.00 (Fig. S3), which indicated that all nodes were strongly supported. In the ML tree, there were two genera (Amomum and Alpinia) in subfamily Alpinioideae, and six genera (Curcuma, Globba, Hedychium, Kaempferia, Stahlianthus, and Zingiber) in subfamily Zingiberoideae with moderate to strong support (bootstrap values = 83–100%) (Fig. 8). Within subfamily Zingiberoideae, there were two clusters Curcuma I and Curcuma II in genus Curcuma: Curcuma II comprised only one species (C. flaviflora), while the rest of the Curcuma species were grouped in Curcuma I; Curcuma I also included genus Stahlianthus; for genus Globba, G. schomburgkii was first grouped with G. schomburgkii var. angustata with strong support (bootstrap value = 100%), and had a nested relationship with G. marantina, G. lancangensis, and G. multiflora (bootstrap values = 100%); for genus Hedychium, H. neocarneum and H. spicatum were first clustered together with strong support (bootstrap value = 90%), and then clustered with H. coccineum and H. coronarium with strong support (bootstrap values = 100%); for genus Kaempferia, there were two clusters with strong support (all bootstrap values = 100%), one of which included two forms of K. rotunda (K. rotunda ‘Red Leaf’ and K. rotunda ‘Silver Diamonds’), and the other included K. galanga and K. elegans; for genus Zingiber, Z. zerumbet, Z. spectabile, Z. montanum, Z. officinale and Z. recurvatum were clustered step-by-step with strong support (bootstrap values ≥ 99%) (Fig. 8). Meanwhile,
### Table 3: Evaluation of the identification capability of thirteen regions among four genera in subfamily Zingiberoideae

| Species                        | Bootstrap values of thirteen regions on ML trees |
|--------------------------------|--------------------------------------------------|
|                               | matK   | ndhF   | ycf1   | trnK-UUU-CD51-rps16-CD52 | psaJ       | ycf4   | cemA   | trnT-UGU-trnL-UAA-CD51 | trnT-GGU-psbD | rpl32-trnL-UAG | psbM-trnD-GUC | ndhF - rpl32 | rps15 - ycf1(1) | ccsA - ndhD |
| G. lancangensis               | 62      | 56     | 66     | 95     | 64                      | 59     | 60     | 71     | 92     | 85                     | 56               | 92           | 93                 |
| G. marantina                  | 99      | 100    | 99     | 92     | 80                      | 100    | 92     | 87     | 98     | 100                    | 99               | 87           | 71                 |
| G. multiflora                 | 99      | 56     | 100    | 99     | 98                      | 82     | 99     | 71     | 92     | 99                     | 99               | 56           | 98                 |
| G. schomburgii                | 89      | 99     | 99     | 92     | 98                      | 72     | 99     | 71     | 92     | 75                     | 91               | 89           | 96                 |
| G. schomburgii var. angustata | 89      | 99     | 99     | 92     | 98                      | 72     | 99     | 71     | 92     | 75                     | 91               | 89           | 96                 |
| H. coccineum                  | 28      | 100    | 50     | 36     | 90                      | 99     | 88     | 73     | 69     | 63                     | 89               | 42           | 88                 |
| H. coronarium                 | 100     | 27     | 100    | 36     | 20                      | 26     | 100    | 100    | 99     | 100                    | 100              | 26           | 20                 |
| H. neocamineum                | 28      | 24     | 82     | 71     | 17                      | 70     | 30     | 54     | 29     | 28                     | 31               | 26           | 14                 |
| H. spicatum                   | 72      | 27     | 82     | 71     | 20                      | 70     | 30     | 73     | 29     | 28                     | 31               | 62           | 20                 |
| K. galanga                    | 40      | 86     | 98     | 58     | 61                      | 72     | 99     | 93     | 33     | 7                      | 94               | 60           | 35                 |
| K. elegans                    | 95      | 86     | 98     | 96     | 61                      | 38     | 99     | 93     | 33     | 22                     | 94               | 53           | 35                 |
| K. rotunda ‘Red Leaf’         | 95      | 100    | 100    | 99     | 97                      | 95     | 87     | 99     | 82     | 97                     | 87               | 93           | 62                 |
| K. rotunda ‘Silver Diamonds’  | 96      | 100    | 100    | 99     | 97                      | 95     | 87     | 99     | 82     | 97                     | 87               | 93           | 62                 |

Note: ratio (%) = [(the total number of species-the number of species with bootstrap values below 50%)/the total number of species] × 100%; **ycf1** here a protein coding gene in chloroplast genome. The sequenced species in this study were marked in bold.
four genera Globba, Hedychium, Kaempferia and Zingiber were monophyletic in the ML tree; Kaempferia and Zingiber were strongly supported as sister genera (bootstrap value = 87%), and Curcuma II was the other sister for Kaempferia with high support (bootstrap value = 100%); Curcuma I and “the Kaempferia clade” consisting of Curcuma II, Kaempferia and Zingiber, were moderately supported as sisters (bootstrap value = 83%). Interestingly, Curcuma I + “the Kaempferia clade”, Hedychium and Globba had a nested evolutionary relationship with strong support (bootstrap values = 99%) in the ML tree (Fig. 8). In conclusion, the BI and ML phylogenetic trees showed clear relationships among the four genera of Globba, Hedychium, Kaempferia and Zingiber in subfamily Zingiberioideae.

### Discussion

Chloroplast genome structure and sequence variation

In this study, 10 complete chloroplast genomes of 9 species from four genera of subfamily Zingiberioideae, namely, Globba, Hedychium, Kaempferia and Zingiber, were sequenced, assembled and applied for their comparative analyses with other related Zingiberioideae species [15, 21, 23, 26]. The genome size (between 162,630 bp and 163,968 bp), genome quadripartite structure (one LSC, one SSC and two IR regions), GC content

| Gene  | Ln L   | Estimates of parameters | Positively selected sites |
|-------|--------|--------------------------|---------------------------|
| accD  | −2726.384456 | p0=0.96351 p1=0.00502 q=0.00512 | 4W 1.000**, 9L 0.987*, 218H 0.958*, 299R 0.968* |
| cctA  | −1712.097311 | p0=0.97744 p1=0.002057 q=0.023055 | 87T 0.957*, 180L 0.994**, 200Y 1.000**, 201K 1.000** |
| ndhA  | −1991.585023 | p0=0.98899 p1=0.01015 q=0.03386 | 132F 0.964*, 1895 1.000**, 1905 1.000**, 191T 1.000**, 192V 1.000** |
| ndhB  | −2119.544926 | p0=0.98391 p1=0.00050 q=0.137861 | 133V 0.953*, 181T 0.957*, 246P 0.955* |
| rpl20 | −162.569004 | p0=0.97242 p1=0.00500 q=0.180939 | 20L 1.000** |
| rbcL  | −2346.044248 | p0=0.97658 p1=0.01215 q=0.20915 | 169L 0.980*, 225I 0.996**, 226Y 0.997**, 247C 0.963*, 255I 0.955*, 407L 0.980*, 424L 0.999**, 4495 1.000** |
| rps10 | −806.763954 | p0=0.94834 p1=0.00500 q=0.01576 | 118K 0.998**, 125Y 1.000** |
| rpsC1 | −3213.853612 | p0=0.98613 p1=0.003585 q=0.990000 | 147N 0.972*, 606D 0.971* |
| rpsC2 | −6806.754011 | p0=0.98887 p1=0.04146 q=0.00956 | 711Y 0.995**, 1174W 0.984* |
| rps12 | −759.324938 | p0=0.73184 p1=0.175806 q=0.00500 | 1M 0.955*, 2P 0.996**, 3T 0.961*, 4I 0.956*, 5K 0.956*, 6Q 1.000**, 7L 0.998**, 8I 0.999**, 9R 0.974*, 10N 0.999**, 11A 0.998**, 12R 0.959*, 13Q 1.000**, 14P 0.966*, 15I 0.989*, 16R 0.959*, 17N 0.999**, 18V 0.999**, 19T 1.000**, 20K 1.000**, 21S 0.998**, 22P 0.998**, 23A 0.963*, 24L 0.998*, 25R 0.986*, 26E 0.998*, 27C 0.964*, 28P 0.998**, 29Q 1.000**, 30R 0.998**, 31R 0.999**, 32G 0.999**, 33T 0.999**, 34C 0.962*, 35T 0.956*, 36R 0.958*, 37V 0.998**, 38p 0.960*, 4495 0.952*, 116Q 0.952* |
| rps18 | −553.014018 | p0=0.98952 p1=0.59674 q=0.990000 | 27P 0.973* |
| ycf1 | −10.584.294185 | p0=0.88238 p1=0.404461 q=0.560436 | 14L 0.994**, 16M 0.985*, 48R 0.961*, 142L 0.990**, 212A 0.989*, 215R 0.981*, 606D 0.975*, 663R 0.944**, 809Y 0.952*, 928P 0.992**, 1293V 0.986*, 1302I 0.964*, 1341M 1.000**, 1433K 0.992**, 1439N 0.999**, 1452K 0.984*, 1453K 0.982*, 1466K 0.999**, 14695 0.998**, 1473D 0.999**, 1499D 0.966*, 1506Q 0.991**, 1528E 0.988*, 1576F 0.999**, 1586Y 1.000**, 1590K 0.990**, 1604P 0.990**, 1621A 0.987*, 1628L 0.991**, 1629N 0.993*, 1632D 0.993*, 1651G 0.987*, 1667S 0.995**, 1757L 0.961* |
| ycf2 | −10.373.971098 | p0=0.93261 p1=0.10353 q=0.15548 | 220P 0.993*, 2340D 0.993*, 1324L 0.994*, 1343F 1.000**, 14115 0.993*, 16651 0.993*, 17503 0.993*, 1757A 0.993*, 212D 0.999*, 2191R 0.994**, 226L 0.963*, 2263H 0.999**, 2265T 0.999*, 2266G 0.995**, 2267E 0.993**, 2268R 0.993*, 2269P 0.999*, 2270P 0.999*, 2272P 0.994* |
| ycf4 | −960.017298 | p0=0.91059 p1=0.00500 q=1.92107 | 181M 0.962*, 184L 0.971* |

Note: the degree of freedom for each gene was 38; * and ** indicate posterior probability higher than 0.95 and 0.99, respectively.
gene composition, most of the protein coding genes, tRNAs and rRNAs showed high similarities among the 10 sequenced chloroplast genomes, which had been observed in other Zingiberoideae chloroplast genomes [15, 21–23, 26]. Although the chloroplast genomes of Zingiberoideae species were highly conserved, gene loss, intron loss and gene duplication occurred in current study, for example, both H. coccineum and H. neocarneum lost lhbA gene, suggesting that gene loss and insertion had occurred during the evolutionary process of H. coccineum and H. neocarneum.

It is important to mention that the chloroplast genome of G. schomburgii lost both trnS-GGA and trnT-GGU, but had two copies of both trnS-GCU and trnT-UGU. Additionally, the trnG-GCC contained one intron in chloroplast genomes of G. lancangensis and G. schomburgii, while the remaining sequenced genomes lost this intron. In comparison to other angiosperms chloroplast genomes, there have been many reports of gene loss of accD, ndh genes, psbE, rpl2, rpl23, trnL-CCA, trnG-GCC, and ycf15, as well as intron loss and gene duplication of rpl2, rpl23, rps15 and ycf1 [36–42].

Contraction and expansion at the borders of the IR regions of chloroplast genomes are considered to be important evolutionary events and may cause size variations, the origination of pseudogenes, gene duplication or the reduction of duplicate genes to single copies [41–44]. For instance, the IR region in Heimia myrtifolia is 25,643 bp long, which is shorter than the IR region of most Lagerstroemia species, indicating that the Lagerstroemia species differentiated later than H. myrtifolia in family Lythraceae [41]. After comparing the chloroplast genomes among 17 Zingiberoideae species, we found that the boundaries between the SSC and two IR regions were relatively conserved, and the distribution and locations of gene types in these regions were highly consistent. Compared with the other 16 Zingiberoideae species, the length of IR region in Z. spectabile was the smallest...
(25,451 bp), mainly because the ycf2 gene located at the LSC/IRB boundary. Additionally, among studied 17 Zingiberoidae species, only two tRNAs of Z. spectabile were found at the LSC/IRB and LSC/IRA boundaries, respectively. Therefore, changes in the LSC/IR boundaries may be the main contributors to the contraction and expansion of IR regions in these Zingiberoidae species.

Highly variable regions can be used as potential DNA barcode markers for the studies on phylogenetic relationships, species identification and population genetics [37, 41, 42, 45]. However, for some Globba, DNA barcode markers have been considered to be difficult to identify. For example, the DNA barcode markers, ITS, matK and trnK-matK, could not discriminate G. atrosanguinea and G. crustisii [2], and G. fecunda and G. multifolia [10]. Meanwhile, based on the Pi values studied here, it is also obvious that the frequently used chloroplast markers, including matK, trnK-matK and trnL-trnF, present low polymorphisms (0.0117, 0.0154, 0.0134, respectively) at subfamily level. Therefore, it is important to explore more highly variable regions at subfamily level that represent potential markers, which can be used for future studies. Based on the results of mVISTA, CGView, nucleotide diversity and ML trees, 12 divergent regions among 17 Zingiberoidae species are suitable for species identification at subfamily and genus level. By comparison, we find that 5 of them are also reported in some Zingiberoidae species, such as ccsA-ndhD, ndhF-rpl32, ycf1 and ycf4-cema reported in Alpinia [13, 19], ycf1 reported in Curcuma [18], ndhF-rpl32 reported in Kaempferia [21], and trnT-UGU-trnL-UAA and ycf1 reported in Zingiber [15, 26]. Our results further verify the reliability and effectiveness of these 5 divergent regions. Among 9 divergent regions, ndhF, ndhF-rpl32, psbM-trnD-GUC, rpl32-trnL-UGA and trnK-UGUU-rps16 are also reported in Lythraceae [45], ndhF and ycf1 are also reported in Monsteracea [46], rpl32-trnL-UGA, rps15-ycf1, trnK-UGUU-rps16, and trnT-UGU-psBD are also reported in Spathiphyllum [46], rpl32-trnL-UGA is also reported in Euterpe [47], rps15-ycf1 is also reported in Prunus [48], and trnK-rps16 (exon2-intron), trnT-trnL and ycf1 are also reported in Allium [49]. Based on these results, we suggest that these divergent regions can be used for potential marker resources of subfamily Zingiberoidae in studies of species identification and phylogeny.

Chloroplast genome evolution in subfamily Zingiberoidae

To resolve the evolutionary history of Zingiberoidae species, it is necessary to analyze their adaptive evolution. The Ka/Ks ratio is very useful for measuring selection pressure at the protein level: if Ka/Ks > 1, the protein is considered to be positively selected; if Ka/Ks = 1, the protein is neutral; and if Ka/Ks < 1, the protein is considered to have undergone purifying selection [50, 51]. In this study, 14 genes with positive selection sites are identified in four genera of Globba, Hedychium, Kaempferia and Zingiber in subfamily Zingiberoidae. Among these genes containing amino acid positive sites, three genes of them encoding ribosome subunit proteins (rpl20, rps12 and rps18), are involved in chloroplast gene expression, which is essential for chloroplast biogenesis and function [52]. rps12 gene exhibits its variations in intragenic exon location and intron content, and has important effects on evolutionary rates and patterns of molecular evolution in fern [53]. Its spicing activity has been reported to impair photosynthesis and perturb development in Arabidopsis [52]. In our analyses, rps12 gene harbors the highest number (40) of positive amino acid sites within 17 Zingiberoidae species, suggesting that the rps12 gene may play important roles in Zingiberoidae species evolution and development.

Moreover, ten genes have also been identified with positive selection sites in current study, namely, accD, ccsA, ndhB, psbH, rbcL, rpoC1, rpoC2, ycf1, ycf2 and ycf4. Recent studies have indicated that these ten genes with positive selection in some angiosperms may be very common phenomena [14, 26, 37, 45–47, 50, 51, 54]. For examples, accD, ndhB, ycf1 and ycf2 have been reported as positive selection in some Zingiberaceae species [14, 26]; accD, ccsA, rbcL, rpoC1, rpoC2, ycf1, ycf2 and ycf4 have also been identified under positive selection in Orchidaceae [37]; accD, ccsA, psbH, rbcL, ycf1 and ycf4 have also been identified under positive selection in Lythraceae [45, 50]; accD, ccsA, rbcL, rpoC2, ycf1 and ycf2 have also been identified under positive selection in Euterpe [47]; accD, rbcL and ycf2, have also been identified under positive selection in Monsteracea [46]; accD, ndhB and ycf2 have also been identified under positive selection in Pterocarpus [51]; and ycf2 has also been identified under positive selection in Pyrus [54]. Additionally, among these ten genes, we find that rbcL, ycf1 and ycf2 genes possess higher number (8, 34, 20, respectively) of positive amino acid sites within 17 Zingiberoidae species. rbcL gene encodes large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco), and ycf1 and ycf2 genes encode unknown function proteins. Both rbcL and ycf1 genes are intensively used for species identification, phylogenetic, phyleogeography, germplasm conservation and innovative utilization of many plants [55–58]. Lastly, there is one gene (ndhA) encoding subunit of NADH-plastoquinone oxidoreductase, which is found under positive selection with 5 amino acid sites. Among our analyzed species of Globba, Hedychium, Kaempferia and Zingiber here, they own diverse pseudostem heights and habitats; for instance, K. galanga spreads flat on ground, living in open areas, while G. schomburgkii, H.
phylogenetic trees have demonstrated some congruence with respect to the rhizome, and the conspicuous parallel orientation of the plane of the distichy of the leafy logical characters of tribe Zingibereae, which has the par-
On the one hand, taxonomic position of Globba requires some discussion. Globba species 
poorly supported [2, 8, 9]. In this study, both BI and ML nuclear data had strongly supported the monophyly of Hedychium, nuclear ITS and chloroplast matK data had very low resolution or were generally lacking among Globba species [10]; and for Hedychium, nuclear ITS and chloroplast matK/trnL-trnF molecular data had strongly supported the monophyly of Hedychium, but its relationships with other genera had been poorly supported [2, 8, 9]. In this study, both BI and ML phylogenetic trees have demonstrated some congruence with previous phylogenies in subfamily Zingiberoideae, for instance, the monophyly of Hedychium [2, 8, 9]. Of course, our analyses also have strongly identified that Globba, Hedychium, and Curcuma 1 + “the Kaempferia clade” consisting of Curcuma II, Kaempferia and Zingiber, display a nested evolutionary relationship with high resolution in subfamily Zingiberoideae (Fig. 8, Fig. S3).

Regarding Globba, Globba was classified into tribe Globbeae, which was one of the two tribes of subfamily Zingiberoideae in previous phylogenies [2, 10]. In our analyses, the close evolutionary relationship between Globba and Hedychium as well as their genetic boundaries have been identified (Fig. 8, Fig. S3). Therefore, the taxonomic position of Globba requires some discussion. On the one hand, Globba owns the universal morphological characters of tribe Zingibereae, which has the parallel orientation of the plane of the distichy of the leafy shoots with respect to the rhizome, and the conspicuous and often well-developed lateral staminodes [2, 3, 10]. On the other hand, Hedychium is classified into tribe Zingibereae, and Globba is close to Hedychium with high resolution based on current phylogenetic results (Fig. 8, Fig. S3). We confirm that our results do not completely resolve all of relationships among genera in two tribes of subfamily Zingiberoideae, and that our results do not sample a great deal of Globba species. In spite of all that, based on the results of our molecular phylogenies, we suggest a realignment of the tribe of Globba in subfamily Zingiberoideae: Globba is here transferred into tribe Zingibereae (Fig. 8, Fig. S3). Because of two important genera, Gagnepainia and Hemiorchis, were classified into tribe Globbeae in past phylogenies [2, 10], we recommend that retaining the tribe Globbeae as previous recognized until future new evidence proves otherwise.

Conclusions
In this study, ten complete chloroplast genomes from nine Zingiberoideae species, namely, G. lancangensis, G. marantina, G. multiflora, G. schomburgkii, G. schomburgkii var. angustata, H. cocineum, H. neocarneum, K. rotunda ‘Red Leaf’, K. rotunda ‘Silver Diamonds’ and Z. recurvatum, have been sequenced, assembled and annotated for the first time. The structural characteristics of these ten chloroplast genomes are shown to be conservative, which are similar to those reported chloroplast genomes of Zingiberoideae species. Meanwhile, comparative analyses of 18 Zingiberoideae chloroplast genomes have generated 12 highly variable regions, which may be used as a potential source of molecular markers for phylogenetic analysis and species identification. Based on whole chloroplast-derived SNP data, phylogenetic relationships among four genera of Globba, Hedychium, Kaempferia and Zingiber in subfamily Zingiberoideae have been clearly resolved. In addition, at level of amino acid sites, 14 genes are under positive selection with high posterior probabilities, which may play important roles in Zingiberoideae species evolution and adaption to diverse environment. These results increase the genomic resources available for subfamily Zingiberoideae, which will be useful for future studies of evolution and phylogenetic.

Methods
Plant material sampling and chloroplast DNA extraction
We generated data on ten chloroplast genomes for nine species within the Zingiberoideae subfamily. Fresh and healthy leaves (G. lancangensis, G. marantina, G. multiflora, G. schomburgkii, G. schomburgkii var. angustata, H. cocineum, H. neocarneum, two forms of K. rotunda, namely K. rotunda ‘Red Leaf’ and K. rotunda ‘Silver Diamonds’, and Z. recurvatum) were collected from the resource garden of the environmental horticulture research institute (23°23′N, 113°26′E) at the Guangdong Academy of Agricultural Sciences, Guangzhou, China, and were immediately frozen in liquid nitrogen and stored at –80 °C. The total genomic DNA was extracted from these leaves using the improved sucrose gradient centrifugation method [59]. The concentration and quantity of each isolated genomic DNA sample were
determined with a NanoDrop 2000 micro spectrometer (Wilmington, DE, USA) and 1% agarose gel electrophoresis, respectively.

**Chloroplast genome sequencing, assembly and annotation**

For *G. schomburgkii*, two libraries with insert sizes of 300bp and 10kb were constructed after purification and then sequenced on an Illumina Hiseq X Ten instrument (Biozeron, Shanghai, China) and a PacBio Sequel platform (Biozero, Shanghai, China), respectively. For the other species (*G. lancangensis*, *G. marantina*, *G. multiflora*, *G. schomburgkii* var. *angustata*, *H. coccineum*, *H. neocarneum*, *K. rotunda* ‘Red Leaf’, *K. rotunda* ‘Silver Diamonds’, and *Z. recurvatum*), a library with insert sizes of 300bp was constructed after purification for each species and then sequenced on an Illumina Hiseq X Ten instrument (Biozeron, Shanghai, China).

The raw data were assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and the adaptor sequences were removed by Trimmomatic v.0.3 [60]. For the final assembled chloroplast genomic sequence of *G. schomburgkii*, the Illumina paired-end clean reads and PacBio clean data were assembled using the previously described method [19, 21, 26], with the software SOAPdenovo v.2.04 [61], Geneious v.11.0.4 [62], BLASR [63], Celera Assembler v.8.0 [64] and GapCloser v.1.12 in SOAPdenovo v.2.04 [61]. The other species’ Illumina clean sequences were assembled into complete chloroplast genomes with SOAPdenovo v.2.04 with default parameters [61] using the chloroplast genome of *G. schomburgkii* as the reference (Table S1).

Complete chloroplast genomes were annotated using the online tool DOGMA (Dual Organellar Genome Annotator) [65] with default parameters, and then, it was checked manually. tRNAs and rRNAs were annotated by BLASTn searches on the NCBI website. A verification of tRNAs and rRNAs was performed using tRNAscanSE with default parameters [66] (Tables 1, 2; Tables S2, S3). Finally, the OGDRAW v.1.3.1 program was used with default settings to draw the circular chloroplast genome maps of the Zingiberidae species and was manually edited [30] (Fig. 1; Fig. S1).

**Prediction of codon usage and RNA editing sites**

To examine the deviation in synonymous codon usage, the relative synonymous codon usage (RSCU) was calculated using MEGA v.7 [32] (Fig. 2a, Table S4). When the RSCU value > 1.00, this means that the use of a codon is more frequent than expected, and vice versa [39, 43]. The clustered heat map of RSCU values of ten sequenced Zingiberidae chloroplast genomes was conducted by R v.3.6.3 [33] (Fig. 2b, Table S5). To predict possible RNA editing sites in the ten sequenced chloroplast genomes, protein coding genes were used to predict potential RNA editing sites using the online program Predictive RNA Editor for Plants (PREP) suite (http://prep.unl.edu/) with a cutoff value of 0.8 [67] (Table S6).

**Analyses of SSRs and long repeats**

MicroSatellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/) was used to detect the location and number of SSRs of the ten sequenced chloroplast genomes with the settings as follows: ≥ 8 for mono-, ≥ 5 for di-, ≥ 4 for tri-, and ≥3 for tetra-, penta-, and hexa-nucleotide SSRs (Fig. 3, Tables S7, S8). The REPuter software was employed to identify long repeats such as forward, palindrome, reverse and complement repeats [68]. The criteria for determining long repeats were as follows: (1) a minimal repeat size of more than 30 bp; (2) a repeat identity of more than 90%; and (3) a hamming distance equal to 3 (Fig. 4, Table S9).

**Comparative genomic analysis**

To detect the contractions and expansions of the IRs in the chloroplast genomes of the four genera (*Globba*, *Hedychium*, *Kaempferia* and *Zingiber*), 18 whole genomes within subfamily Zingiberidae were compared (Fig. 5). Using the annotated chloroplast genome of *G. lancangensis* as the reference, the mVISTA tool with the Shuffle-LAGAN mode [34] was used to make pairwise alignments among these 18 whole chloroplast genomes (Fig. 6). Additionally, the *G. lancangensis* chloroplast genome was compared to 17 chloroplast genomes of the four genera (*Globba*, *Hedychium*, *Kaempferia* and *Zingiber*) using CGView [31] (Fig. 1). GC distributions were measured based on GC skew using the following equation: GC skew = (G-C)/(G+C). The nucleotide variability (Pi) of the protein coding, intron and intergenic regions among these 18 chloroplast genomes within the Zingiberidae subfamily was extracted and then calculated using DnaSP v.6 [69] (Fig. 7, Table S10). To identify the highly divergent regions, the protein coding regions and intergenic regions were extracted and then manually aligned using ClustalW in MEGA v.7 [32]. To obtain the molecular markers for differentiating the four genera in subfamily Zingiberidae, a maximum likelihood (ML) analysis based on the nucleotide substitution model of Tamura-Nei was conducted using MEGA v.7 with 1000 replicates [32] (Table 3, Fig. S2).
were constructed using an SNP matrix from 56 chloroplast genomes by comparing pairwise [70]. A total of 11,835 Ka/Ks were obtained; the value could not be calculated if Ks = 0 or if the two aligned sequences existed as a perfect 100% match (Table S11). Next, to identify amino acid sites under the true occurrence of positive selection, program CODEML from PAML package v.4.8a [35] was used, with each corresponding protein coding sequence of chloroplast genome of *G. lancangensis* as the reference. A Bayes empirical bayes (BEB) approach [71] was then used to calculate posterior probabilities that a site came from the site class with $\alpha > 1$ by a site specific model 8 ($\beta$ and $\omega$). In the BEB analysis, posterior probability higher than 0.95 and 0.99 indicated sites that were under positive selection and strong positive selection, respectively (Table 4, Table S12).

### Phylogenetic tree analyses

To reconstruct the phylogenetic relationships and analyze the phylogenetic positions of 5 *Globba* species, 4 *Hedychium* species, 3 *Kaempferia* species and 5 *Zingiber* species within subfamily Zingiberoideae, 56 complete chloroplast genomes of family Zingiberaceae were used for analysis (Table S13). A total of 49 complete chloroplast genomes were downloaded from the GenBank database. *C. pulverulentus*, *C. viridis* and *C. indica* were used as the outgroups. The phylogenetic trees were constructed using an SNP matrix from 56 chloroplast genomes of family Zingiberaceae and 3 chloroplast genomes of outgroups, using the chloroplast genome of *G. lancangensis* as the reference. Reliable SNP sites were obtained using a previously described method that utilized MUMmer and BLAT software [19, 21, 26, 72–74]. For each chloroplast genome, all SNP sites were connected in the same order to obtain a sequence in FASTA format. Each connected FASTA format sequence contained 11,346 SNP markers (Table S14). Multiple FASTA format sequence alignments were performed with ClustalW in MEGA v.7 [32]. The ML analysis of MEGA v.7 was used for the reconstruction of the ML phylogenetic tree with default settings including 1000 bootstrap replications along with the nucleotide substitution model of Tamura-Nei [32] (Fig. 8). Bootstrap values were categorized as strong (> 85%), moderate (70–85%), weak (50–70%), or poor (< 50%) [2]. Lastly, BI analysis was performed using MrBayes v.3.2 [75], using the substitution model GTR with running parameters: the Markov Chain Monte Carlo algorithm was applied for 2 million generations with four Markov chains and sampled of trees every 100 generations, then the first 10% of trees were discarded as burn-in. The software iTOL v.3.4.3 ([http://itol.embl.de/itol.cgi](http://itol.embl.de/itol.cgi)) was used to edit and visualize the final BI tree (Fig. S3).

### Supplementary information

The online version contains supplemental material available at [https://doi.org/10.1186/s12870-021-03315-9](https://doi.org/10.1186/s12870-021-03315-9).
of Ka, Ks and Ka/Ks in 17 Zingiberoideae species with Ka/Ks > 1.00 and p values < 0.05.

Additional file 12: Table S12. Details of positively selected amino acid loci of 14 protein coding genes in 17 Zingiberoideae species computed by PAML [35].

Additional file 13: Table S13. The 59 chloroplast genomes used for phylogenetic analysis.

Additional file 14: Table S14. The 59 chloroplast genomes’ SNP matrix data.

Additional file 15: Figure S11. Gene maps of the other 9 assembled Zingiberoideae chloroplast genomes in this study. Genes shown inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise. The gray arrowheads indicate the direction of the genes. Different genes are color coded. The innermost darker gray corresponds to GC content, whereas the lighter gray corresponds to AT content. The inner circle also indicates that the chloroplast genome contains a large single copy region (LSC), a small single copy region (SSC) and two copies of the inverted repeat (IRA and IRB).

Additional file 16: Figure S12. Maximum likelihood (ML) trees of five Globba species, four Hedychium species, three Kaempferia species and five Zingiber species based on the chloroplast genome divergent genes and intergenic regions. A ML tree based on the sequences of gene matK. B ML tree based on the sequences of gene ndhF. C ML tree based on the sequences of gene ycf1. D ML tree based on the intergenic sequences of trnK-UGU-psbA-trnT-UGU. E ML tree based on the intergenic sequences of trnH-psbA-trnL-UAA-CDS1. G ML tree based on the intergenic sequences of trnG-UUC-psbD. I ML tree based on the intergenic sequences of rpl32-trnL-UAG. J ML tree based on the intergenic regions of psbH-trnD-GUC. K ML tree based on the intergenic sequences of ndhF-psbJ. L ML tree based on the intergenic sequences of rps15-ycf1. M ML tree based on the intergenic sequences of cssA-ndhD.

Additional file 17: Figure S13. Molecular phylogenetic tree based on the SNPs from 56 chloroplast genomes of family Zingiberaeae using Bayesian inference. The numbers at the nodes were Bayesian inference posterior probabilities. All nodes of the tree were supported by 1.00 Bayesian inference posterior probability. The branch length was proportional to the inferred divergence level and the scale bar indicated the number of inferred nucleic acids substitutions per site. C indica, C. pulvulentus and C. viridis were used as the outgroups.

Acknowledgements
We are grateful to the Guangdong basic and applied basic foundation project, the creative agricultural research team, the research and development project in key areas of Guangdong province and the special financial fund of the Foshan-Guangdong Agricultural Science and Technology Demonstration City Project in 2019. The funding agencies had no role in the design of the study, analysis, interpretation of the data and writing of the manuscript.

Availability of data and materials
All the ten sequenced complete chloroplast genomes in this study have been submitted to NCBI (https://www.ncbi.nlm.nih.gov) with accession numbers MQ62735, and MT473704–MT473712 (see Table 1). These data will remain private until the related manuscript has been accepted. Other chloroplast genomes for phylogenetic analysis can be obtained from NCBI, and their accession numbers are listed in Table S13. All other data generated in this manuscript are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflicts of interest.

Received: 19 November 2020 Accepted: 3 November 2021

Published online: 23 November 2021

References
1. Wu D, Larsen K. Zingiberaceae. Flora of China, vol. 24. Beijing: Science press; 2000. p. 322–77.
2. Kress WJ, Prince LM, Williams KJ. The phylogeny and a new classification of the gingers (Zingiberaceae) evidence from molecular data. Am J Bot. 2002;89:1682–96.
3. Brannen TM. Hardy ginger: including Hedychium, Roscoea and Zingiber. Portland and London, Timber press; 2005. p. 15–242.
4. Gao JY, Xia YM, Huang JY, Li QI, Zhongguo Jiangke Huahui. 1st ed. Beijing: Science press; 2006. p. 40–121.
5. Wu D, Liu N, Ye Y. The Zingiberaceous resources in China. 1st ed. Wuhan: Huazhong university of science and technology university press, 2016.
6. AI TM, Zhang SR, Yang XW, Du LJ, Yan ZY. Zhongguo Yaoqing Zhiwu. In: Dai LK, Zheng YN, Du GH LYC, editors. Zhongguo Yaoyong Zhiwuzhi, vol. 12. Beijing: Peking university medical press; 2013. p. 376–415.
7. Droungchawee J, Kulising C, Sukeaw S, Na Pombejra S, Chavasiri W, Plabutong N, et al. Volatile chemical composition, antibacterial and anti-fungal activities of extracts from different parts of Globba schomburgkii Hook.F. Chem Biodivers. 2019;16(5):e1900057.
8. Ngamriabsakul S, Newman MF, Cronk QCB. The phylogeny of tribe Zingiberaceae (Zingiberaceae) based on its (nrDNA) and trnL-trnF (cpDNA) sequences. Edinb J Bot. 2004;60(3):483–507.
9. Wood TH, Whitten WM, Williams NH. Phylogeny of Hedychium and related genera (Zingiberaceae) based on ITS sequence data. Edinb J Bot. 2000;57(2):261–70.
10. Williams KJ, Kress WJ, Manos PS. The phylogeny, evolution, and classification of the genus Globba and tribe Globbeae (Zingiberaceae): appendages do matter. Am J Bot. 2004;91(1):100–14.
11. Techaprasan J, Klinbunga S, Ngamriabsakul S, Jenjitikul T. Genetic variation of Kaempferia (Zingiberaceae) in Thailand based on chloroplast DNA (psbA-trnH and petA-psbA) sequences. Genet Mol Res. 2010;9:1957–73.
12. Wu M, Li Q, Hu Z, Li X, Chen S. The complete Anomum kravanh chloroplast genome sequence and phylogenetic analysis of the commelindas. Molecules. 2017;22:1875.
13. Cui Y, Chen X, Nie L, Sun W, Hu H, Lin Y, et al. Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible Anomum species. Int J Mol Sci. 2019;20:4040.
14. Gao B, Yuan L, Tang T, Hou J, Pan K, Wei N. The complete chloroplast genome sequence of Alpinia oxyphylla Miq. And comparison analysis within the Zingiberaceae family. PLoS One. 2019;14:e0218817.

15. Cui Y, Nie L, Sun W, Xu Z, Wang Y, Yu J et al. Comparative and phylogenetic analyses of geng (Zingiber officinale) in the family Zingiberaceae based on the complete chloroplast genome. Plants (Basel). 2019:283.

16. Zhang Y, Deng J, Li Y, Gao G, Ding C, Zhang L et al. The complete chloroplast genome sequence of Curcuma flaviflora (Curcuma). Mitochondrial DNA A. 2016;27:3644–5.

17. Gui L, Jiang S, Xie D, Yu L, Huang Y, Zhang Z et al. Analysis of complete chloroplast genomes of Curcuma and the contribution to phylogeny and adaptive evolution. Gene. 2020;722:143455.

18. Liang H, Zhang Y, Deng J, Gao G, Ding C, Zhang L et al. The complete chloroplast genome sequences of 14 Curcuma species: insights into genome evolution and phylogenetic relationships within Zingiberaceae. Front Genet. 2020;11:802.

19. Li DM, Zhu GF, Xu YC, Ye YJ, Liu JM. Complete chloroplast genomes of three medicinal Alpinia species: genome organization, comparative analyses and phylogenetic relationships in family Zingiberaceae. Plants (Basel). 2020;9:286.

20. Li DM, Zhao CY, Zhu GF, Xu YC. Complete chloroplast genome sequence of Arumium villorum. Mitochondrial DNA B. 2019:9:2673–4.

21. Li DM, Zhao CY, Liu XF. Complete chloroplast genome sequences of Kaempferia galanga and Kaempferia elegans: molecular structures and comparative analysis. Molecules. 2019;24:474.

22. Li DM, Xu YC, Zhu GF. Complete chloroplast genome of the plant Stahlianthus involucatus (Zingiberaceae). Mitochondrial DNA B. 2018;9:2702–3.

23. Li DM, Zhao CY, Zhu GF, Xu YC. Complete chloroplast genome sequence of Hedyctium coronarium. Mitochondrial DNA B. 2018;9:2806–7.

24. Li DM, Zhao CY, Xu YC. Characterization and phylogenetic analysis of the complete chloroplast genome of Curcuma longa (Zingiberaceae). Mitochondrial DNA B. 2019:9:2974–5.

25. Li DM, Zhu GF, Xu YC, Ye YJ, Liu JM. Characterization and phylogenetic analysis of the complete chloroplast genome of Curcuma zedoaria (Zingiberaceae). Mitochondrial DNA B. 2020;5:1329–31.

26. Li DM, Ye YJ, Xu YC, Liu JM, Zhu GF. Complete chloroplast genomes of Zingiber montanum and Zingiber zerumbet: genome structure, comparative and phylogenetic analyses. PLoS One. 2020;15:e0236590.

27. Wicke S, Schneeweiß GM, DePamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76:273–97.

28. Brunkard JO, Runkel AM, Zambryski PC. Chloroplast extend stromules independently and in response to internal redox signals. Proc Natl Acad Sci U S A. 2015;112(32):10044–9.

29. Greiner S, Lehewak P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47(W1):W59–64.

30. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17:134.

31. Greiner S, Lehewak P, Ruck K. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47(W1):W59–64.

32. R Core Team. R: a language and environment for statistical computing. https://www.R-project.org. Accessed 20 May 2020.

33. Kumar S, Stecher G, Tamura K. MEGA XV: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.

34. R Core Team. R: a language and environment for statistical computing. https://www.R-project.org. Accessed 20 May 2020.

35. Frailey DC, Chalvadkar SR, Vaughn JG, Cooney CG, Benetzen JL. Gene loss and genome rearrangement in the plastids of five Hemiparasites in the family Orobanchaceae. BMC Plant Biol. 2018;18(1):30.

36. Liu E, Yang C, Liu J, Jin S, Harjiani N, Hu Z et al. Comparative analysis of complete chloroplast genome sequences of four major Amorphophallus species. Sci Rep. 2019;9:8909.

37. Yan X, Liu T, Yuan X, Xu Y, Yan H, Hao G. Chloroplast genomes and comparative analyses among thirteen taxa within Mynsinaceae. Str. Clade (Mynsinioideae, Primulaceae). Int J Mol Sci. 2019:20:4534.

38. Zheng G, Wei L, Ma L, Wu Z, Gu C, Chen K. Comparative analyses of chloroplast genomes from 13 Lagerstroemia (Lythraceae) species: identification of highly divergent regions and inference of phylogenetic relationships. Plant Mol Biol. 2020;102(6):659–76.

39. Henriquez CL, Abdullah AI, Carlens MM, Zuluaga A, Croat TB, McKain MR. Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). Genomics. 2020;12:2349–60.

40. Munyao JD, Dong X, Yang JX, Mbandi EM, Wanga VO, Oulo MA et al. Complete chloroplast genomes of Chlorophyton comosum and Chlorophyton galliforme: genome structures, comparative and phylogenetic analysis. Plants (Basel). 2020;9:296.

41. Abdullah MF, Shahzadi I, Waseem S, Mirza B, Ahmed I, Waheed MT. Chloroplast genome of Hibiscus rosa-sinensis (Malvaceae): comparative analyses and identification of mutational hotspots. Genomics. 2020;11(2):581–91.

42. Gu C, Ma L, Wu Z, Chen K, Wang Y. Comparative analyses of chloroplast genomes from 22 Lythraceae species: inferences for phylogenetic relationships and genome evolution within Myrtales. BMC Plant Biol. 2019;19:281.

43. Henriquez CL, Abdullah AI, Carlens MM, Zuluaga A, Croat TB, McKain MR. Comparative analysis of the complete chloroplast genome of Curcuma aromatica (Zingiberaceae): genome structures, comparative and phylogenetic analyses. Plants (Basel). 2020;9:2172.

44. de Santana Lopes A, Pacheco TG, da Silva ON, do Nascimento Vieira L, Guerra MR, Mattar EPL et al. Plastid genome evolution in Amazonian acai palm (Euterpe oleracea Mart.) and Atlantic forest acai palm (Euterpe edulis Mart.). Plant Mol Biol 2021;105(4–5):559–574.

45. Xue S, Shi T, Luo W, Ni X, Iqbal S, Ni Z et al. Comparative analysis of the complete chloroplast genome among Punica mume, P. ameraica, and P. salicina. Hortic Res. 2019:6:89.

46. Huo YM, Gao LM, Liu BJ, Yang YJ, Kong SP, Sun YQ et al. Complete chloroplast genome sequences of four Allium species: comparative and phylogenetic analyses. Sci Rep. 2019:9:12250.

47. Yan M, Zhao X, Zhou J, Huo Y, Ding Y, Yuan Z. The complete chloroplast genomes of Punica granatum and a comparison with other species in Lythraceae. Int J Mol Sci. 2019;20:2886.

48. Hong Z, Wu Z, Zhao K, Yang Z, Zhang Z, Zhang N et al. Comparative analyses of five complete chloroplast genomes from the genus Pentacarpus (Fabaceae). Int J Mol Sci. 2020;21:5758.

49. Lee K, Leister D, Kleine T. Arabidopsis mitochondrial transcription termination factor mTERF2 promotes splicing of group II introns. Cells. 2021;10(2):315.

50. Liu S, Wang Z, Wang H, Su Y, Wang T. Patterns and rates of plastid rps12 gene evolution inferred in a phylogenetic context using plastidomic data of ferns. Sci Rep. 2020;10(19):9394.

51. Li Y, Zhang J, Li L, Gao L, Xu J, Yang M. Structural and comparative analysis of the complete chloroplast genome of Pyrus hopeiensis—“wild plants with a tiny population”—and three other Pyrus species. Int J Mol Sci. 2018;19:3262.

52. Ho VT, Tran TKP, Vu TT, Widiarsih S. Comparison of matK and rbcL DNA barcodes for genetic classification of jewel orchid accessions in Vietnam. J Genet Eng Biotechnol. 2021;19(1):93.

53. Li H, Xiao W, Tong T, Li Y, Zhang M, Lin X et al. The specific DNA barcodes based on chloroplast genes for species identification of Orchidaceae plants. Sci Rep. 2021;11:1424.

54. El-Banhawy A, Nour HH, Acedo C, Elkholy A, Fairied A, Al-Juhani W et al. Taxonomic revisiting and phylogenetic placement of two endangered plant species: Silene leucophylla Boiss. and Silene schimpanseniana Boiss. (Caryophyllaceae). Plants (Basel). 2021;10(4):740.

55. Li WW, Liu LQ, Zhang QP, Zhou WQ, Fan GQ, Liao K et al. Comparative analysis of Prunus armeniaca L. revealed by chloroplast DNA and nuclear ribosomal sequences. Sci Rep. 2021;11(1):13623.

56. Li X, Hu Z, Lin L, Xu Q, Gao H, Luo G et al. High-throughput pyrosequencing of the complete chloroplast genome of Magnolia officinalis and its application in species identification. Acta Pharm Sin. 2012;47:124–30.
60. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexibletrimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
61. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-end de novo assembler. Gigascience. 2012;1:14.
62. Kearse M, Moir R, Wilson A, Stoneshavas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9.
63. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinform. 2012;13:238.
64. Denisov G, Wielenz B, Halpen AL, Miller J, Axerlod N, Levy S, et al. Consensus generation and variant detection by celeria assembler. Bioinformatics. 2008;24:1035–40.
65. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20:3252–5.
66. Lowe TM, Chan PP. tRNAscan-SE On-line: Search and contextual analysis of transfer RNA genes. Nucleic Acids Res. 2016;44:W54–7.
67. Mower JP. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 2009;37:W253–9.
68. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–42.
69. Rozas J, Ferrer-Mateu A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large datasets. Mol Biol Evol. 2017;34(12):3299–302.
70. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinform. 2010;8:77–80.
71. Yang Z, Wong WSW, Nielsen R. Bayes empirical bayes inference of amino acids sites under positive selection. Mol Biol Evol. 2005;22(4):1107–18.
72. Marcia G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome assembly. PLoS Comput Biol. 2018;14:e1005944.
73. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
74. Bhagwat M, Young L, Robison RR. Using BLAT to find sequence similarity in closely related genomes. Curr Protoc Bioinformatics. 2012;Chapter 10(Unit10.8). https://doi.org/10.1002/0471250953.bi1008s37.
75. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–42.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.