BMC Biotechnology

RESEARCH

Open Access

# Chloroplast genome sequencing based on genome skimming for identification of Eriobotryae Folium

Fang Li[1,2†], Xuena Xie[2†], Rong Huang[2], Enwei Tian[2], Chan Li[2] and Zhi Chao[2,3*]

## Abstract

**Background:** Whole chloroplast genome (cpDNA) sequence is becoming widely used in the phylogenetic studies of plant and species identification, but in most cases the cpDNA were acquired from silica gel dried fresh leaves. So far few reports have been available to describe cpDNA acquisition from crude drugs derived from plant materials, the DNA of which usually was seriously damaged during their processing. In this study, we retrieved cpDNA from the commonly used crude drug Eriobotryae Folium (*Pipaye* in Chinese, which is the dried leaves of *Eriobotrya japonica*, PPY) using genome skimming technique.

**Results:** We successfully recovered cpDNA sequences and rDNA sequences from the crude drug PPY, and bioinformatics analysis showed a high overall consistency between the cpDNA obtained from the crude drugs and fresh samples. In the ML tree, each species formed distinct monophyletic clades based on cpDNA sequence data, while the phylogenetic relationships between *Eriobotrya* species were poorly resolved based on ITS and ITS2.

**Conclusion:** Our results demonstrate that both cpDNA and ITS/ITS2 are effective for identifying PPY and its counterfeits derived from distantly related species (i.e. *Dillenia turbinata* and *Magnolia grandiflora*), but cpDNA is more effective for distinguishing the counterfeits derived from the close relatives of *Eriobotrya japonica*, suggesting the potential of genome skimming for retrieving cpDNA from crude drugs used in Traditional Chinese Medicine for their identification.

**Keywords:** *Eriobotrya japonica*, Eriobotryae Folium, Crude drug, Identification, Chloroplast genome, Genome skimming

## Background

Chloroplast is one of the two organelles having their own genetic materials in plant cells. The chloroplast genomes (cpDNA) are double-stranded DNA in a closed-loop configuration with a length ranging from 120 to 220 kb [1–3]. The cpDNAs, which are maternally inherited and remain haploidy without recombination, have multiple copies per cell and in angiosperms, their size, structure and gene composition are quite consistent [4–7]. The cpDNA contains rich genetic information, based on which a large database can be constructed for comparative study. In addition, the moderate nucleotide substitution rate of cpDNAs and the differences in their molecular evolution speed of the coding region and non-coding region allow for systematic studies of the plants at different levels [8–11]. The good collinearity of the cpDNAs of different plant groups also provides much convenience for comparative analysis and can reflect the phylogenetic history of the plant population [12–15].

*Correspondence: chaozhi@smu.edu.cn
†Fang Li and Xuena Xie contributed equally to this work
[2] Faculty of Medicinal Plants and Pharmacognosy, School of Traditional Chinese Medicine, Southern Medical University, Guangzhou 510515, China
Full list of author information is available at the end of the article

The development of high-throughput sequencing technology has allowed full-length sequencing of the cpDNA [16, 17], which has become an important basis of phylogenomic studies. The complete sequence of cpDNA has confirmed some non-genomic-data-based conclusions at different classification levels and revealed many new systematic relationships; it has also shown unique advantages in species identification [12, 13, 18–21]. Using massively parallel sequencing technology, Nock et al. [22] sequenced the cpDNA of *Oryza sativa japonica* and two other *Oryza* species (i.e. *O. meridionalis* and *O. australiensis*), together with that of *Potamophila parviflora* (a close relative to *Oryza*) and *Microlaena stipoides* (an Australian native grass), and found that each species could be identified accurately based on these cpDNA sequences. In the following years, increasing reports emerged on the application of cpDNA sequences in the identification of such medicinal plants as *Magnolia officinalis* [23], *M. grandiflora* [24], *Scutellaria baicalensis* [25], *Fritillaria cirrhosa* [26], and *Ligularia spp.* [27]. According to incomplete statistics, the cpDNA of at least 3721 plant species have been described so far, ranging from green algae to terrestrial and aquatic plants [28].

In almost all these studies, fresh leaves were used as the samples for acquiring cpDNA. No report has been available to describe cpDNA sequencing using samples of crude drugs derived from medicinal plants, the DNA of which was usually damaged during preparation [29, 30]. To investigate the feasibility of cpDNA sequencing based on samples of crude drugs, we attempted to obtain complete chloroplast genome through genome skimming from crude drugs derived from different parts (root, rhizome, fruit and seed) from Pipaye (PPY), the dried leaf of loquat [*Eriobotrya japonica* (Thunb.) Lindl.], the selected representative of leaf-derived crude drugs.

In Traditional Chinese Medicine, PPY is believed to be effective for treating asthma and coughing [31]. Nin Jiom Pei Pa Koa, a Chinese patent medicine with loquat leaf as the main ingredient, has attracted aroused heated discussion in the United States during the influenza season in 2018 after the Wall Street Journal published a report portraying an architect and professor of design at Pratt Institute for taking the medicine to cure his long-standing cough [32]. Actually, the history of using PPY for medical purposes can be dated back to Han Dynasty [33]. In the long history of its medicinal uses, PPY is sometimes confused with the leaves of some other plants, e.g. *Dillenia turbinata* and *Magnolia grandiflora*, which are similar in appearance to loquat leaves [34]. These counterfeits have no effects of genuine PPY, thus should be clearly identified, but their identification can be difficult even for professionals due to their high similarities in appearance, especially when the leaves are cut into pieces.

Theoretically, the Internal transcribed spacer region (ITS) can be used for loquat species identification, but currently no studies of ITS-based identification of PPY against its adulterants has been reported, except for some studies on genetic diversity of *Eriobotrya japonica* [35, 36]; nor was a specific PCR system has been available for PPY identification. Currently, a thin-layer chromatography (TLC) inspection for PPY is recommended in the Chinese Pharmacopoeia, in which ursolic acid serves as the reference substance. As ursolic acid is widely distributed in plant species, the TLC-based identification of crude drugs often has a low specificity. Although a UPLC-Q-TOF/MS analysis targeting the anti-EGFR chemical constituents had been reported for PPY identification [37], the performance of this modality for PPY identification remains to be further verified.

cpDNA sequencing is a promising technique for crude drug identification. Genome skimming is PCR-free to avoid such issues of amplification failure and false positive and false negative results. With genome skimming, not only the cpDNA sequence but also the sequence of ITS region can be obtained from the high-throughput sequencing data, thus a combined analysis of cpDNA and ITS sequences can be possible. Additionally, genome skimming is more cost-effective than MALDI-TOF MS analysis.

In this study, we sequenced the cpDNA not only from fresh leaf samples of *Eriobotrya japonica* and its close relatives *E. deflexa*, *E. cavaleriei*, *E. fragrans*, as well as those of *Dillenia turbinata* and *Magnolia grandiflora*, but also from self-made sun-dried *E. japonica* leaves (self-prepared PPY, SP) and three commercial PPY samples to investigate the feasibility of cpDNA sequencing in identification of the crude drugs. We also compared the efficiency of cpDNA sequencing and the general barcode such as ITS/ITS2 for PPY identification.

## Results

### Analysis of cpDNAs of *Eriobotrya japonica* and its relative and counterfeit species

#### *Structure and genes*

In this study, all the cpDNAs showed a typical circular tetramerous structure, consisting of a pair of inverted repeats (IRs), a large single copy region (LSC), and a small single copy region (SSC) (Fig. 1). The size of cpDNA and its regions were all similar across different *Eriobotrya* species (Table 1). The cpDNA length of genus *Eriobotrya* ranges from 159,115 bp (*E. japonica*) to 159,393 (*E. deflexa*); the cpDNA length is 159,270 bp for *E. cavaleriei* and 159,177 bp for *E. fragrans*. The size of the IR region ranges from 26,317(*E. fragrans*) to 26,335 bp (*E. cavaleriei*), while the SSC and LSC size varies from 19,213 (*E. fragrans*) to 19,350 bp (*E. cavaleriei*)

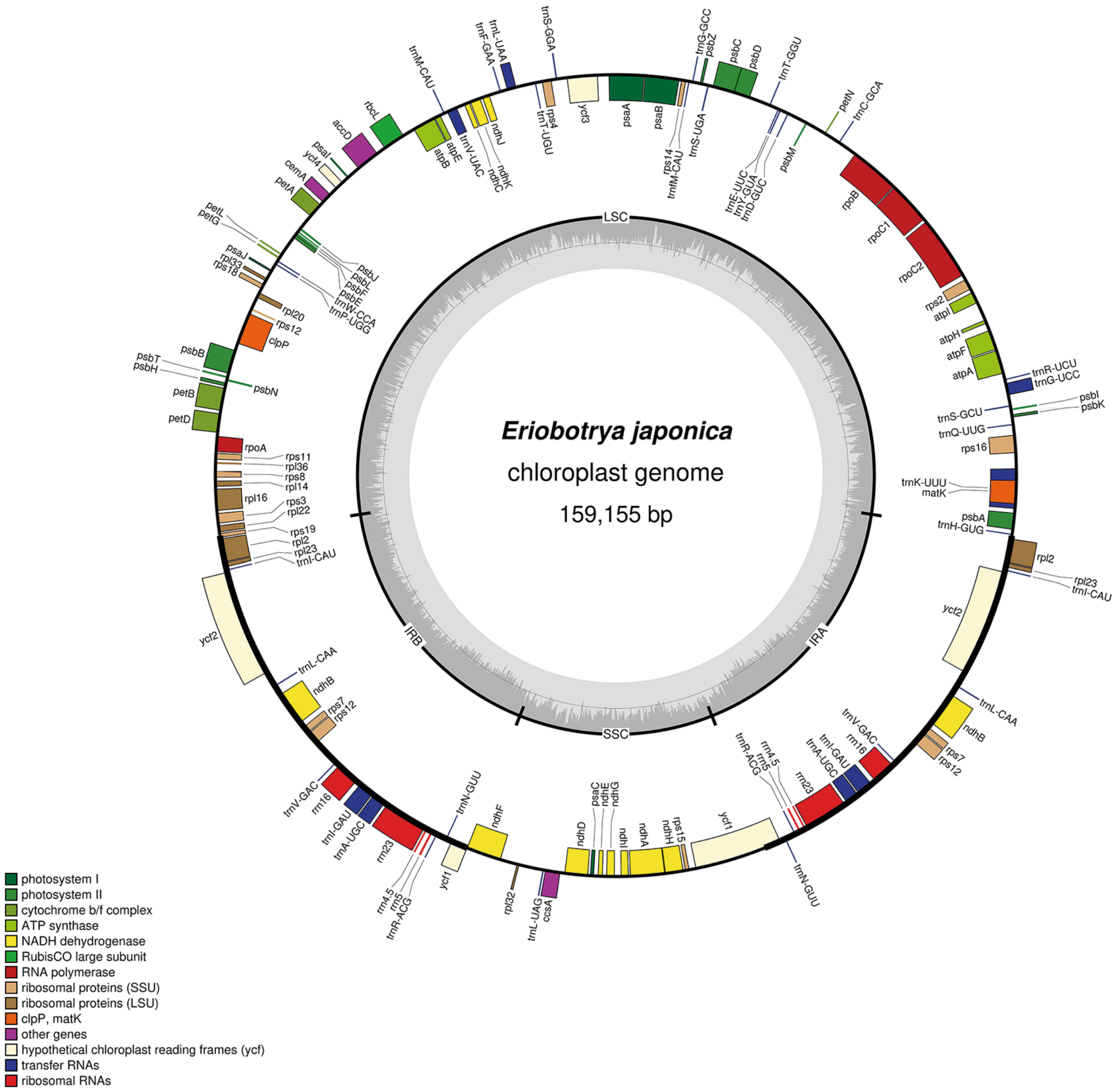

**Fig. 1** Chloroplast genome map of *E. japonica*. The genes outside of the circle are transcribed clockwise, while those inside are transcribed counterclockwise. Small single copy (SSC), large single copy (LSC), and inverted repeats (IRa, IRb) are indicated

and from 87,222 (*E. japonica*) to 87,401 bp (*E. deflexa*), respectively (Table 1). The cp gonomes of *D. turbinata* and *M. grandiflora* are 163,250–159,690 bp in length, consisting of an IR region of 26,497–26,580 bp, a SSC region of 18,754–19,349 bp and LSC regions of 87,776–90,907 bp. *E. japonica* contains 112 genes, including 78 protein coding genes, 30 tRNA genes and 4 rRNA genes, the same as the remaining *Eriobotrya* species and *M. grandiflora*. The *D. turbinata* cpDNA consists of 113 genes, including 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. Compared to the *Eriobotrya* species, *D. turbinata* has 113 genes due to the presence of the gene *infA*. In addition, the presence of *infA* and the deletion of *rpl22* gene of *M. grandiflora* result in the consistency in the number of genes with *Eriobotrya* species. The *ycf1* sequence located in the IRa and SSC boundary of all the samples was identified as a pseudogene because it was truncated, i.e. incomplete duplications of the normal copy. In addition, two pseudogenes, *accD* and *ndhK*, were also found in *D. turbinata*. In the cpDNA of all the

**Table 1** Summary of cpDNA characteristics of 11 samples

| | ***E. japonica*-1** | ***E. japonica*-2** | **PPY-1** | **PPY-2** | **PPY-3** | **SP** | ***E. cavaleriei*** | ***E. deflexa*** | ***E. fragrans*** | ***D. turbinata*** | ***M. grandiflora*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DNA concentration (ng/μL) | 273.1 | 310.1 | 41.9 | 29.4 | 33.6 | 9.76 | 88 | 21.9 | 47.5 | 43.8 | 362.7 |
| Coverage | 309.71 ± 45.78 | 747.45 ± 99.19 | 60.02 ± 20.91 | 140.14 ± 28.56 | 188.80 ± 35.90 | 59.99 ± 20.96 | 505.66 ± 44.15 | 579.12 ± 94.80 | 492.26 ± 54.56 | 359.61 ± 71.08 | 258.53 ± 36.78 |
| Size (bp) | 159,115 | 159,156 | 159,155 | 159,156 | 159,156 | 159,202 | 159,270 | 159,393 | 159,177 | 163,250 | 159,690 |
| Number of genes | 112 | 112 | 112 | 112 | 112 | 112 | 112 | 112 | 112 | 113 | 112 |
| Number of protien-coding genes | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 79 | 78 |
| Number of tRNA genes | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |
| Number of rRNA genes | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| Overall GC content | 36.70% | 36.70% | 36.70% | 36.70% | 36.70% | 36.70% | 36.70% | 36.70% | 31.00% | 36.10% | 39.3% |
| LSC length (bp) | 87,222 | 87,223 | 87,222 | 87,223 | 87,223 | 87,271 | 87,250 | 87,401 | 87,330 | 90,907 | 87,776 |
| IR length (bp) | 26,326 | 26,326 | 26,326 | 26,326 | 26,326 | 26,325 | 26,335 | 26,330 | 26,317 | 26,497 | 26,580 |
| SSC length (bp) | 19,281 | 19,281 | 19,281 | 19,281 | 19,281 | 19,281 | 19,350 | 19,332 | 19,213 | 19,349 | 18,754 |
| GC content in IR (%) | 42.70% | 42.70% | 42.70% | 42.70% | 42.70% | 42.70% | 42.70% | 42.70% | 42.70% | 42.80% | 42.70% |
| Reference sequence | KY085905, MN577877, NC034639 for cpDNA<br>MH711704, MG938044, MF096288, KX675082 for ITS/ITS2 | | | | | | MN577877 for cpDNA KJ170784, KP093136 for ITS/ITS2 | MK920282 for cpDNA MG938042, JQ392434 for ITS/ITS2 | MN577877 for cpDNA MH246945, KP093137 for ITS/ITS2 | NC042740 for cpDNA AY096031 for ITS/ITS2 | JN867587, NC020318, MN990594 for cpDNA |

samples, the gene *rps12* was a trans-splicing gene, whose 5' exon was located in the LSC region and the 3' exon in the IRs region.

The junction positions were conserved in *Eriobotrya* species. *Eriobotrya* species have partially duplicated *rps19* and *ndhF* genes in the IR regions, while these two genes are located respectively in the LSC and SSC regions of *D. turbinata* and *M. grandiflora* (Fig. 2). In *Eriobotrya* species, the extent of *rpsl9* duplication ranges from 120 (*E. cavaleriei* and *E. fragrans*) to 127 bp (*E. deflexa*), and 12 nucleotides of *ndhF* are duplicated. The final 12 nucleotides of the IR region are shared by *ndhF* and the pseudogene *ycf1* (*ψycf1*), which are transcribed in opposite directions; in *D. turbinata* and *M. grandiflora*, *ψycf1* gene is located in the IRb region and *ndhF* in the SSC region. The LSC/IRa-*rpl2* spacer ranges in length between 38 (*D. turbinata*) and 195 nucleotides (*E. deflexa*).

### *Condon usage*

The total length of the protein coding genes (PCGs) of *Eriobotrya* cpDNAs ranges from 78,600 (*E. fragrans*) to 78,630 bp (*E. cavaleriei* and *E. deflexa*), and that of *D. turbinata* and *M. grandiflora* was 77,301 bp and 77,811 bp, respectively (Table 2). These PCGs contain 25,767 (in *D. turbinata*) to 26,210 (in *E. cavaleriei* and *E. deflexa*) codons, with UGA, UAG and UAA as the termination codons. For *Eriobotrya* cpDNAs, the most frequent amino acid is leucine (Leu), encoded by 2749–2754 (10.51%) of the codons; the least frequent amino acid in the cpDNAs is cysteine (Cys), encoded by 299–301 (1.14%) of the codons (Fig. 3). Most of the amino acid codons have preferences except for methionine and tryptophan. Within the PCGs of *Eriobotrya* cpDNAs, the GC content of the codons in the third position was 26.7%. Within the PCGs of *D. turbinata* and *M. grandiflora* cpDNAs, the AT content of the codons at the third position is 26.4% and 28.8%, respectively. All the preferred synonymous codons (RSCU > 1) of *E. japonica* ended with A or U except for the codons of *trnL-CAA*, while most of the non-preferred synonymous codons (RSCU < 1) ended with G or C, which is the same as the other *Eriobotrya* species in our study.

### *SSRs and long repeat sequences*

We found that the mononucleotide repeats of genus *Eriobotrya*, *D. turbinata* and *M. grandiflora* were by far the most frequent SSR type, followed by dinucleotides, tetranucleotides, trinucleotides, pentanucleotides, and finally hexanucleotide (Table 3). *Eriobotrya* cpDNAs exhibit variations in the number of SSRs; the number is 92 in *E. japonica*, 90 in *E. cavaleriei*, 108 in *E. deflexa* and 98 in *E. fragrans*. The number of SSRs is 93 in *D. turbinata*, and is only 53 in *M. grandiflora*, the smallest among all the species. among the *Eriobotrya* species, there was no trinucleotide repeat and only a single hexanucleotides was found only in *E. deflexa*. No pentanucleotide repeat was found in *M. grandiflora*.

The tandem repeats in the cpDNAs of *Eriobotrya* species has generally a low variation, ranging from 130 (*E. fragrans*) to 133 (*E. cavaleriei*) (Table 4). Among all the species, *D. turbinata* has the highest number of tandem repeats (up to 216), while *M. grandiflora* has the least number of only 49. Five different long repeats, including tandem, complement, forward, palindromic and reverse repeats, were found in the cpDNA in this study. Complement repeat was absent in *E. japonica, E. fragrans* and *M. grandiflora*. Reverse repeat was not found in *M. grandiflora*.

### *Highly divergent regions*

In the cpDNA of each species, the non-coding regions have a greater variability than the coding regions (Fig. 4). Several divergent regions such as *trnH-GUG* , *petN-psbM*, and *trnT-GGU-psbD* were found in *Eriobotrya* species. For all the species, some highly variable regions were observed in the intergenic regions, as in *trnH-GUG*, *trnK-UUU-rps16*, *petN-psbM*, *trnT-GGU-trnL-UAA*, *rpl20-rps12*, *psbZ-trnG-GCC* (Fig. 5). The *ndhF-rpl32* region showed the highest average sequence divergence (0.1126), followed by *rpl32-trnL-UAG* (0.1202), *rps16-trnQ-UUG* (0.11007), and *accD-psbI* (0.1076) (Fig. 5), with the remaining genes having a divergence less than 0.1.

### Comparison of cpDNAs obtained from PPY, SP and *E. japonica* fresh leaves

The average cover of fresh samples of *E. japonica* (309.71–747.45) was as high as about 5 times that of the dried samples (59.99–188.80). Both PPY and SP were consistent with *E. japonica* in terms of gene number, GC content (Table 1), genetic makeup (Table 5), the boundaries of IR region (Fig. 2), codon usage (Table 2), and SSRs type and number (Table 3). Both of PPY and SP had 112 genes with a GC content of 36.7%, including 78 protein coding genes, 30 tRNA genes and 4 rRNA genes. In structural analysis of cpDNAs, only minor variations were observed in terms of the length of cpDNAs (from 159,115 bp in *E. japonica* to 159,202 bp in SP) (Table 1) and the amount of long repeat sequences (Table 4). SP had one more forward repeats and two more tandem repeats than *E. japonica*, while PPY was similar with *E. japonica* in the amount and type of the long repeat sequences.

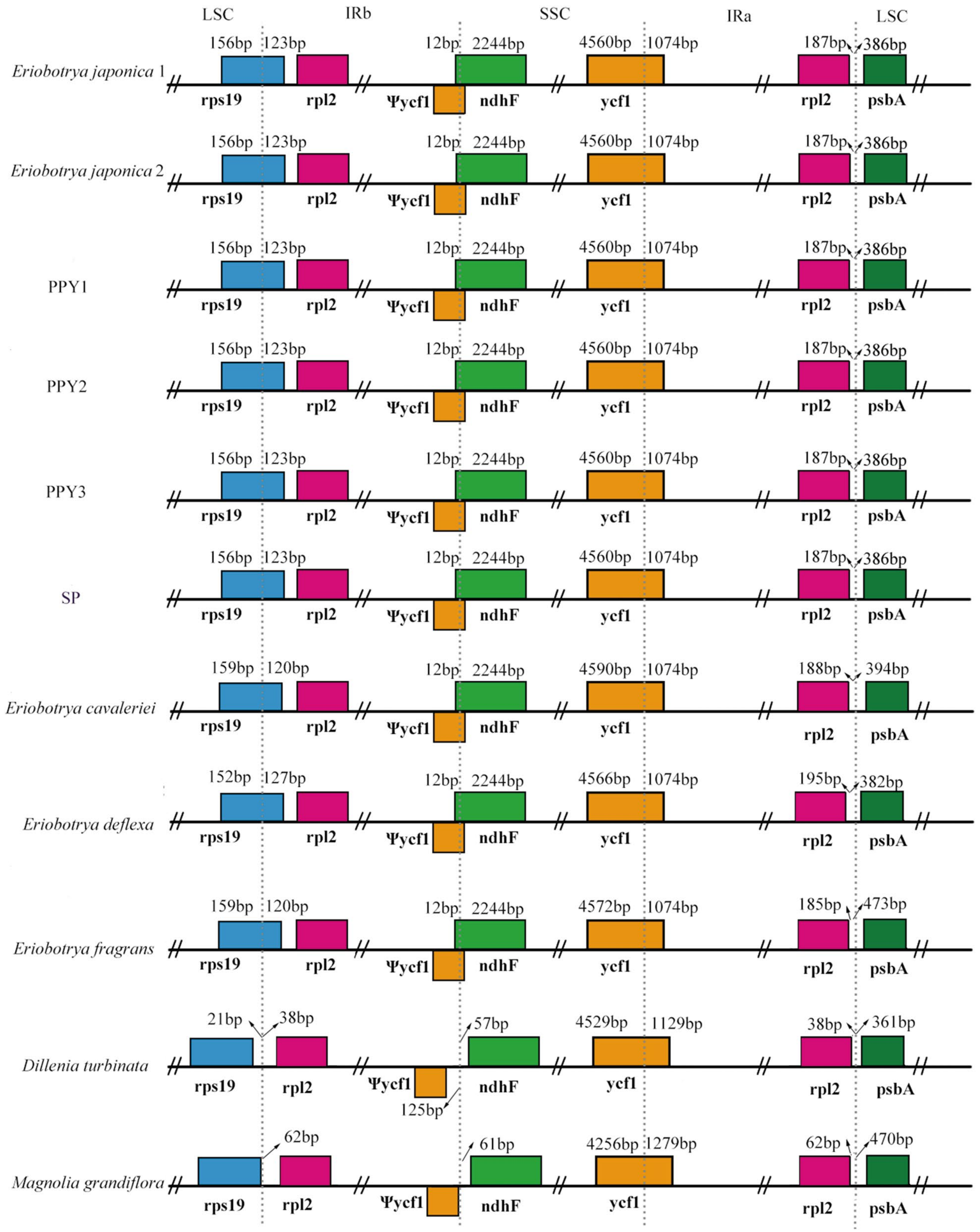

**Fig. 2** Comparison of the border regions of the LSC, SSC and IR regions among the 11 cp genomes. The genes cross the LSC/IRb or IRb/SSC regions, indicating that the LSC/IRb boundary has moved backward or the IRb/SSC boundary moves forward in these species

**Table 2** Indexes of codon usage bias in 11 samples representing 6 species

| | ***E. japonica*-1** | ***E. japonica*-2** | **PPY-1** | **PPY-2** | **PPY-3** | **SP** | ***E. cavaleriei*** | ***E. deflexa*** | ***E. fragrans*** | ***D. turbinata*** | ***M. grandiflora*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PCG length (bp) | 78,612 | 78,612 | 78,612 | 78,612 | 78,612 | 78,612 | 78,630 | 78,630 | 78,600 | 77,301 | 77,811 |
| Codon Number | 26,204 | 26,204 | 26,204 | 26,204 | 26,204 | 26,204 | 26,210 | 26,210 | 26,200 | 25,767 | 26,122 |
| Amino acid No | 26,120 | 26,120 | 26,120 | 26,120 | 26,120 | 26,120 | 26,126 | 26,126 | 26,116 | 25,682 | 26,038 |
| SC No | 25,052 | 25,052 | 25,052 | 25,052 | 25,052 | 25,052 | 25,058 | 25,058 | 25,049 | 24,632 | 24,959 |
| ENC | 49.52 | 49.52 | 49.52 | 49.52 | 49.52 | 49.52 | 49.51 | 49.52 | 49.52 | 49.33 | 50.66 |
| CAI | 0.166 | 0.166 | 0.166 | 0.166 | 0.166 | 0.166 | 0.166 | 0.166 | 0.165 | 0.165 | 0.167 |
| CBI | −0.106 | −0.106 | −0.106 | −0.106 | −0.106 | −0.106 | −0.106 | −0.106 | −0.107 | −0.114 | −0.096 |
| FOP | 0.350 | 0.350 | 0.350 | 0.350 | 0.350 | 0.350 | 0.350 | 0.350 | 0.350 | 0.346 | 0.357 |
| GC content (%) | 0.378 | 0.378 | 0.378 | 0.378 | 0.378 | 0.378 | 0.378 | 0.378 | 0.378 | 0.375 | 0.392 |
| $GC_3$ content (%) | 0.267 | 0.267 | 0.267 | 0.267 | 0.267 | 0.267 | 0.267 | 0.267 | 0.267 | 0.264 | 0.288 |

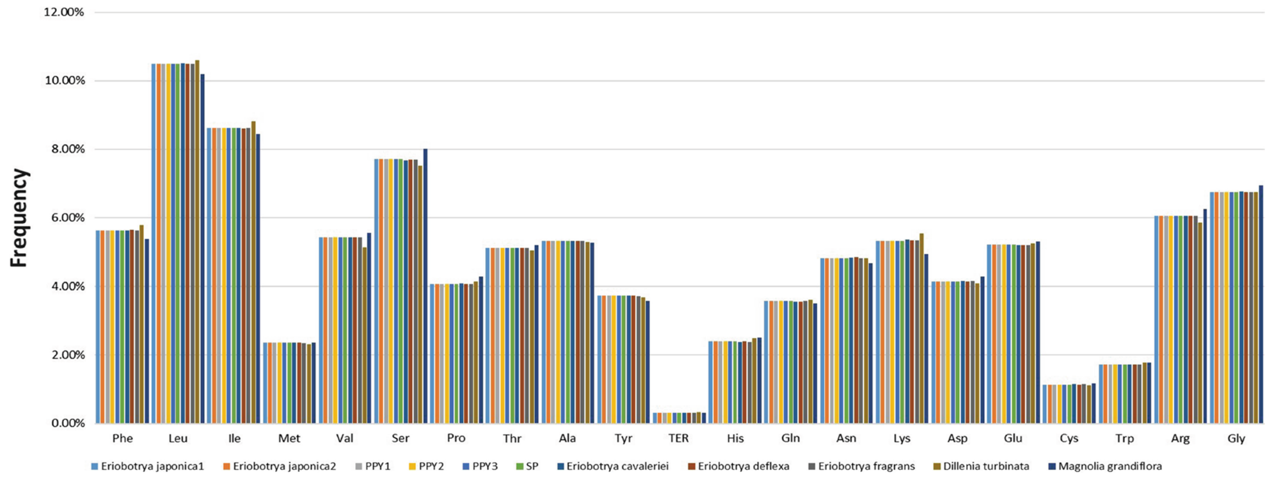

**Fig. 3** Amino acid frequencies in 11 samples protein-coding sequences

**Table 3** Comparison of simple repeats (SSR) in 11 cp genomes

| Sample | Mononucleotides | Dinucleotides | Trinucleotides | Tetranucleotides | Pentanucleotides | Hexanucleotides | Total |
|---|---|---|---|---|---|---|---|
| *E. japonica*-1 | 70 | 15 | 0 | 6 | 1 | 0 | 92 |
| *E. japonica*-2 | 70 | 15 | 0 | 6 | 1 | 0 | 92 |
| PPY-1 | 70 | 15 | 0 | 6 | 1 | 0 | 92 |
| PPY-2 | 70 | 15 | 0 | 6 | 1 | 0 | 92 |
| PPY-3 | 70 | 15 | 0 | 6 | 1 | 0 | 92 |
| SP | 70 | 15 | 0 | 6 | 1 | 0 | 92 |
| *E. cavaleriei* | 70 | 15 | 0 | 4 | 1 | 0 | 90 |
| *E. deflexa* | 83 | 17 | 0 | 6 | 1 | 1 | 108 |
| *E. fragrans* | 75 | 17 | 0 | 6 | 0 | 0 | 98 |
| *D. turbinata* | 58 | 18 | 6 | 9 | 1 | 1 | 93 |
| *M. grandiflora* | 30 | 9 | 3 | 9 | 0 | 2 | 53 |
| Total | 736 | 166 | 9 | 70 | 9 | 4 | 994 |
| Ratio | 74.04% | 16.70% | 0.91% | 7.04% | 0.91% | 0.40% | 100% |

**Table 4** Comparison of long repeat sequences in 11 cp genomes

| Sample | Tandem repeat | Complement repeat | Forward repeat | Palindromic repeat | Reverse repeat | Total |
|---|---|---|---|---|---|---|
| *E. japonica*-1 | 131 | 0 | 25 | 20 | 3 | 179 |
| *japonica*-2 | 131 | 0 | 25 | 20 | 3 | 179 |
| PPY-1 | 131 | 0 | 25 | 20 | 3 | 179 |
| PPY-2 | 131 | 0 | 25 | 20 | 3 | 179 |
| PPY-3 | 131 | 0 | 25 | 20 | 3 | 179 |
| SP | 133 | 0 | 26 | 20 | 3 | 182 |
| *E. cavaleriei* | 133 | 2 | 23 | 18 | 7 | 183 |
| *E. deflexa* | 132 | 1 | 22 | 16 | 11 | 182 |
| *E. fragrans* | 130 | 0 | 22 | 17 | 11 | 180 |
| *D. turbinata* | 216 | 1 | 19 | 19 | 8 | 263 |
| *M. grandiflora* | 49 | 0 | 11 | 16 | 0 | 76 |
| total | 1448 | 4 | 248 | 206 | 55 | 1961 |

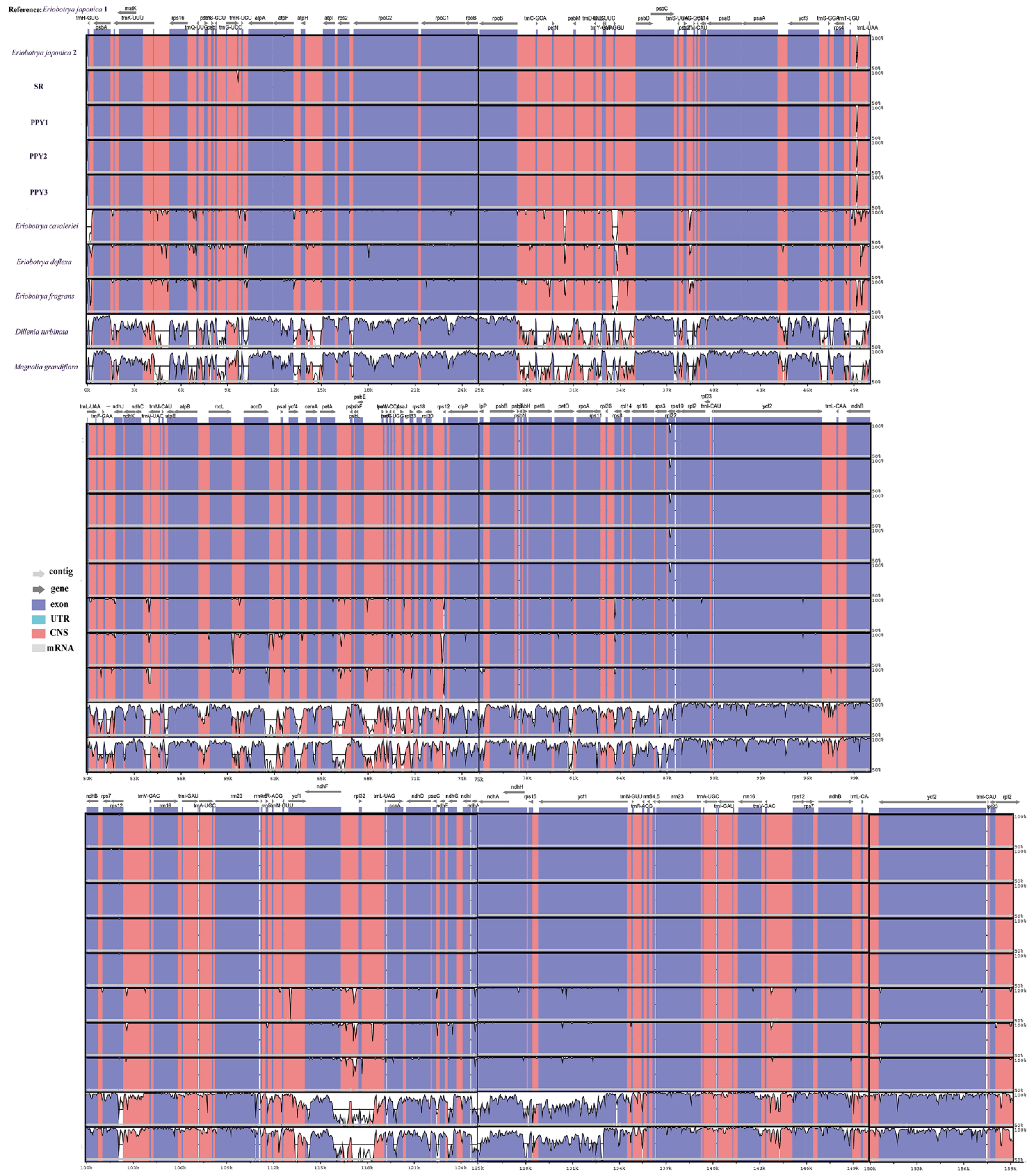

**Fig. 4** Comparative chloroplast genomic analysis. The red area represents the non-coding area, and the purple area represents the coding area. The large twists and turns indicate large variations

## Phylogenetic tree and species identification

Among all the species, the topological structures of ITS, ITS2 and cpDNAs were basically identical, including three major clades, namely *Eriobotrya*, *Dillenia* and *Magnolia* species (Figs. 6, 7, Additional file 1: Fig. S1). But the phylogenetic positions based on ITS and ITS2 of the other *Eriobotrya* species were different in that *E. cavaleriei* was placed close to *E. deflexa* or *E. fragrans*

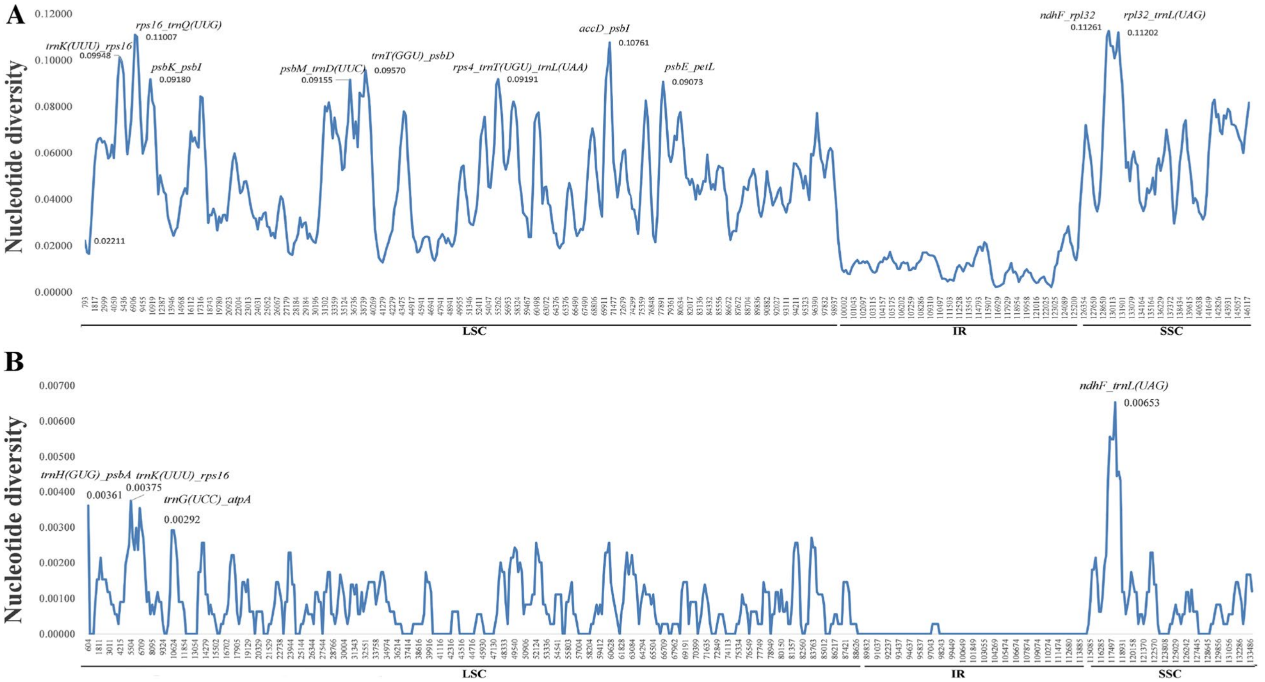

**Fig. 5** Comparative analysis of the nucleotide diversity (Pi) value of the cp genomes among the 11 species. **A** Coding regions, **B** non-coding regions

**Table 5** List of genes found in *Eriobotrya japonica* cpDNA

| Gene category | Gene group | Gene name |
|---|---|---|
| Photosynthesis related genes | Photosystem I | *psaA, psaB, psaC, psaI, psaJ* |
| | photosystem II | *psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ* |
| | Cytochrome b/f complex | *petA*(× 2), *petB*, petD*, petG, petL, petN* |
| | ATP synthase | *atpA, atpB, atpE, atpF, atpH, atpI* |
| | NADH dehydrogenase | *ndhA, ndhB**(× 2), *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK* |
| | RubisCO large subunit | *rbcl* |
| Transcription and translation related genes | Ribosomal proteins(SSU) | *rps2, rps3, rps4, rps7*(× 2), *rps8, rps11, rps12***(× 2), *rps14, rps15, rps16*, rps18, rps19* |
| | Ribosomal proteins(LSU) | *rpl2**(× 2), *rpl14, rpl16*, rpl20, rpl22, rpl23*(× 2), *rpl32, rpl33, rpl36* |
| RNA genes | Ribosomal RNAs | *rrn4.5*(× 2), *rrn5*(× 2), *rrn16*(× 2), *rrn23*(× 2) |
| | Transfer RNAs | *trnS-GGA, trnS-UGA, trnS-GCU, trnE-UUC, trnT-UGU, trnT-GGU, trnF-GAA, trnM-CAU, trnW-CCA, trnP-UGG, trnI-CAU*(× 2), *trnI-GAU**(× 2), *trnL-CAA*(× 2), *trnL-UAA*, trnL-UAG, trnV-GAC*(× 2), *trnV-UAC*, trnR-ACG(*× 2), *trnR-UCU, trnN-GUU*(× 2), *trnH-GUG, trnQ-UUG, trnC-GCA, trnD-GUC, trnY-GUA, trnG-UCC*, trnfM-CAU, trnK-UUU*, trnA-UGC**(× 2), *trnG-GCC* |
| | RNA polymerase | *ropA, ropB, ropC1*, ropC2* |
| Other genes | | *ccsA, accD, cemA, clpP**, matK* |
| Proteins of unknown function | ycf | *ycf1, ycf2*(× 2), *ycf3**, ycf4* |

with strong support (Fig. 7; Additional file 1: Fig. S1). In addition, the *Dillenia* species was closely related to *Eriobotrya* species, as shown in Fig. 7. The ML tree based on cpDNA had a higher resolution and each genus node had a bootstrap value of 100% (Fig. 6). PPY, SP and *E. japonica* were all classified into one clade with a bootstrap value of 100%.

Based on the K2P model, the intraspecific genetic distances ranged from 0.0005 (*E. japonica*) to 0.0889 (*E. cavaleriei*), from 0.0026 (*E. japonica*) to 0.1403 (*E.*

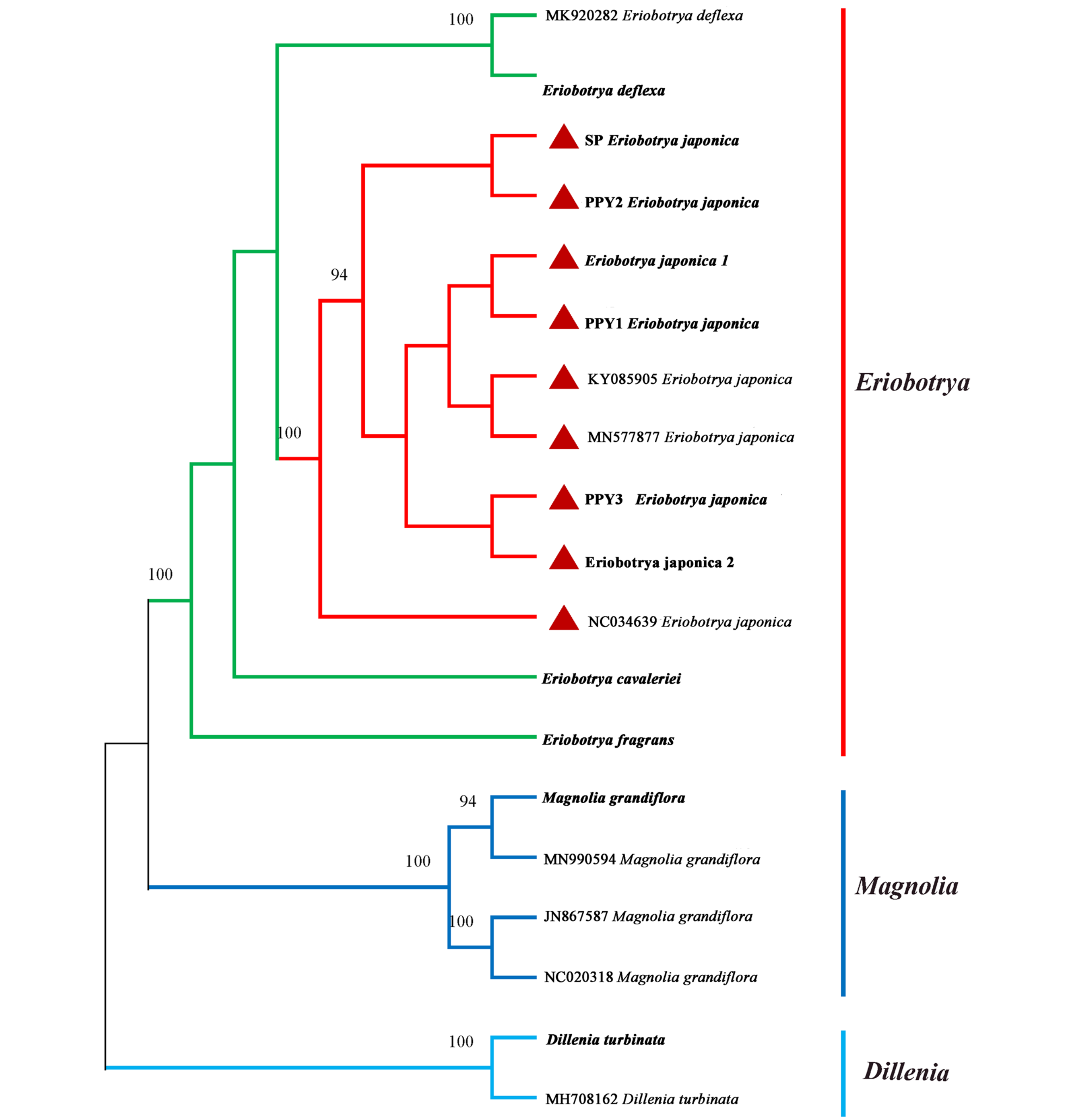

**Fig. 6** Phylogenetic tree constructed using ML based on complete cp genomes. The number above the branches are bootstrap support values

*fragrans*), and from 0.0000 (*E. japonica*) to 0.0004 (*M. grandiflora*) in the cases of ITS, ITS2, and cpDNA, respectively; the interspecific genetic distances ranged from 0.0285 (*E. japonica* and *E. deflexa*) to 0.8665 (*M. grandiflora* and *D. turbinata*), from 0.0371 (*E. japonica* and *E. deflexa*) to 0.7495 (*M. grandiflora* and *E. fragrans*), and from 0.0007 (*E. japonica* and *E. deflexa*) to 0.1195 ((*M. grandiflora* and *D.indica*), respectively.

## Discussion

The cpDNA of higher plants is highly conserved, which ensures the direct homology of genes among distant evolutionary groups. Compared with nuclear and mitochondrion genome, cpDNA has a greater gene density with a moderate evolution rate, thus making cpDNA a suitable and unique molecule for accurate species identification. Currently few studies have been available to report plant

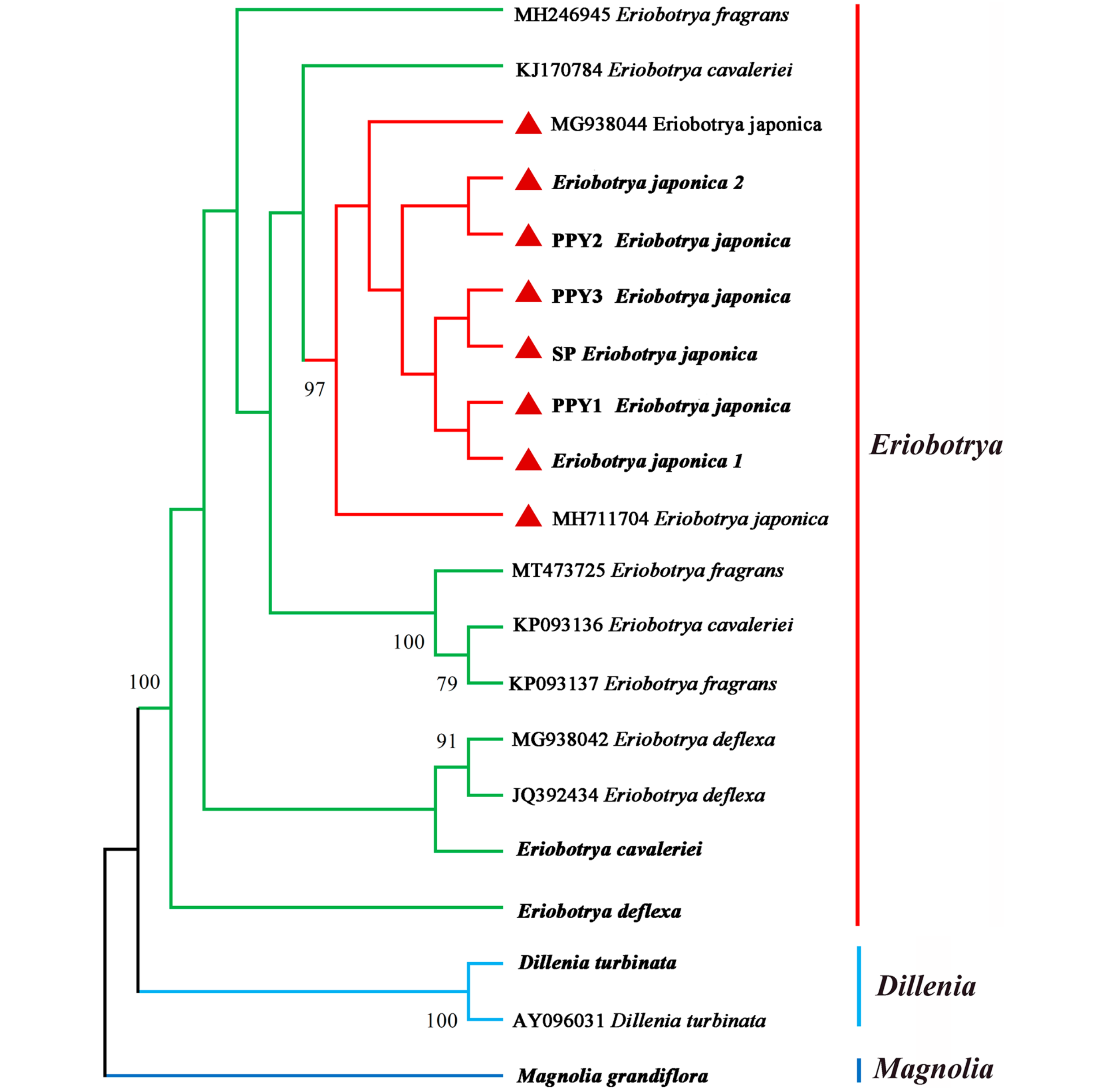
MH246945 *Eriobotrya fragrans*
KJ170784 *Eriobotrya cavaleriei*
MG938044 Eriobotrya japonica
*Eriobotrya japonica 2*
PPY2 *Eriobotrya japonica*
PPY3 *Eriobotrya japonica*
SP *Eriobotrya japonica*
PPY1 *Eriobotrya japonica*
*Eriobotrya japonica 1*
MH711704 *Eriobotrya japonica*
MT473725 *Eriobotrya fragrans*
KP093136 *Eriobotrya cavaleriei*
KP093137 *Eriobotrya fragrans*
MG938042 *Eriobotrya deflexa*
JQ392434 *Eriobotrya deflexa*
*Eriobotrya cavaleriei*
*Eriobotrya deflexa*
*Dillenia turbinata*
AY096031 *Dillenia turbinata*
*Magnolia grandiflora*
97
100
79
100
91
100
*Eriobotrya*
*Dillenia*
*Magnolia*

**Fig. 7** Phylogenetic tree constructed using ML tree based on 20 ITS sequences. The numbers above the branches are bootstrap support values

species identification by sequencing cpDNA from crude drugs derived from plants instead of fresh leaves. To test the feasibility of acquiring complete cpDNA through genome skimming for crude drug identification, we used commercial PPY samples purchased from local pharmacies, i.e. the crude drug practically sold to patients, not merely silica gel dried fresh leaf materials used in previous studies. To our best knowledge, such a pilot empirical study has not been reported previously.

Different from that in silica gel dried fresh leaf materials, the genomic DNA in crude drugs usually have severe degradation, as often seen in the specimens stored for a long time. Long storage time can result in DNA degradation [30] and DNA fragmentation [29] to cause difficulties in the genome sequencing and identification. Genome skimming has proved to well suit the needs of species identification based on degenerated genome DNA, and researchers have successfully sequenced cpDNA from herbarium materials stored for decades with this technique [38–40], which is even capable of sequencing complete or almost complete cpDNA from specimens stored up to 146 years.

As expected, the genomic DNA extracted from the crude drugs was of a poor quality in this study. But with genome skimming, the cpDNAs retrieved were almost identical to those obtained from the fresh samples, and a low amount of degraded genomic DNA (9 ng) was sufficient for operation. cpDNAs acquired from PPY, SP and *E. japonica* samples showed negligible variations, which can be inferred from the same coding genes, tRNAs and rRNAs among their cp genomes. Besides cpDNA, we also successfully recovered rDNA sequences from the crude drug PPY. These results further demonstrate that genome skimming is less affected by template quality than other sequencing methods [38–41].

In the continuous efforts for searching ideal DNA barcodes for plants, ITS/ITS2 have been considered as the most promising ones [42, 43] for their high resolution of inter- and intraspecific relationship [44–47], but so far a widely accepted universal DNA barcode has not been available yet. Appropriate barcodes for specific plant taxonomic groups should be investigated case by case. Theoretically, ITS/ITS2 can be used for *Eriobotrya* species identification with better convenience and at a lower cost compared to cp genome method. Nevertheless, our results confirmed that both cpDNA and ITS/ITS2 were efficient for identifying PPY and its simple counterfeits (*Dillenia turbinata* and *Magnolia grandiflora*), but ITS/ITS2-based identification had a poor resolution for *Eriobotrya* species, *E. japonica* and its close relatives (*E. deflexa, E. cavaleriei, E. fragrans*). Previous studies proposed that the unresolved relationships among them may be attributed to the confusion of the interspecific boundaries between *E. cavaleriei* and *E. fragrans* based on short sequences [48–50]. Overlaps between the intraspecific and interspecific K2P distances based on ITS/ITS2 were also reported. Thus, the short sequences (i.e. rDNA ITS/ITS2) are not as powerful as expected in identifying Eriobotryae Folium and its counterfeits due to insufficient variation information.

CpDNA contains much more genetic information and can provide a large database for species identification [12, 51–53] to significantly increase the resolution at lower taxonomic ranks such as genus and species, and thus may serve as a super barcode for species identification [26], as in the case of *Eriobotrya*. Our phylogenetic analysis based on cpDNA data showed that the samples belonging to the same species formed a separate clade, each with a high bootstrap value. In addition, the intraspecific K2P distance values were significantly lower than the interspecific K2P distance when using cpDNA data. These results demonstrate that, compared to ITS and ITS2 sequences, cpDNA is more effective for the identification of Eriobotryae Folium.

Although cpDNA genome can provide more characteristics and increase the amount of sequence data to enhance species discrimination, it does not address the basic challenge that cpDNA do not necessarily track species boundaries [54]. Substantial numbers of unlinked nuclear markers (e. g. transcriptome sequencing and RAD-seq) should be taken to access the ultimate big gains in the discriminatory power [54].

## Conclusions

Despite of severe degradation of the genomic DNA, cpDNA and rDNA can be successfully sequenced and assembled from crude drug of Eriobotryae Folium through genome skimming. Chloroplast genome sequence data can be more effective than rDNA ITS and ITS2 sequences for the identification of Eriobotryae Folium and the counterfeits with a close resemblance. The results of this study demonstrate that genome skimming is capable of retrieving whole chloroplast genome from crude drugs used in traditional Chinese medicine for their identification.

## Methods

### Plant and crude drug samples

Two samples of fresh leaves of *E. japonica* (*E. japonica*-1 and *E. japonica*-2) were collected from the Medicinal Plant Garden of Southern Medical University and South China Botanical Garden. The fresh leaves of *E. cavaleriei*, *E. deflexa*, *E. fragrans*, *D. turbinata* and *M. grandiflora* were collected from different localities. A portion of the sample *E. japonica*-1 was subjected to sun-drying to prepare self-made PPY sample (SP). Three batches of PPY crude drug (PPY-1, PPY-2 and PPY-3) were purchased from Kangmei Pharmaceutical Co., Ltd, Dongfang Pharmacy, and Henglu Pharmacy, respectively. The voucher specimens and crude drug samples were all identified by the corresponding author (Table 6). The crude drug samples were kept in a cool and dry place, while the fresh leaf samples were kept at − 80 °C.

### DNA extraction

Genomic DNA was extracted from the above samples using the modified CTAB method [55]. To eliminate the interference by phenolic substances on DNA extraction, 20 mg polyvinyl pyrrolidone was mixed with *Eriobotrya* samples before DNA extraction [56]. DNA concentration and quality were examined using a NanoDrop 2000C spectrophotometer and by 1.2% agarose gel electrophoresis.

**Table 6** Information of samples

| Samples | Collecting site locality | Geographical coordinates | Specimen voucher/batch no. | GenBank accession of cp genome |
|---|---|---|---|---|
| *Eriobotrya japonica*-1 | Medicinal Plant Garden of Southern Medical University | 23° 19′ 45″ N, 113° 34′ 37″ E | Chao Zhi EJ201403 | MT479167 |
| *E. japonica*-2 | South China Botanical Garden | 23° 19′ 23″ N, 113° 37′ 18″ E | Chao Zhi EJ201910 | MT473726 |
| *E. cavaleriei* | Wuhan Botanical Garden | 30° 54′ 49″ N, 114° 43′ 30″ E | Chao Zhi 201,812 | MT473722 |
| *E. deflexa* | Guangdong Tree Park | 23° 20′ 13″ N, 113° 38′ 05″ E | Chao Zhi ED201812 | MT473724 |
| *E. fragrans* | Chenhedong Nature Reserve, Guangdong | 23° 44′ 02″ N, 113° 50′ 64″ E | Chao Zhi EF201903 | MT473725 |
| *Dillenia turbinata* | South China Botanical Garden | 23° 18′ 51″ N, 113° 36′ 77″ E | Chao Zhi DT201403 | MT473723 |
| *Magnolia grandiflora* | Medicinal Plant Garden of Southern Medical University | 23° 19′ 45″ N, 113° 34′ 37″ E | Chao Zhi MG201403 | MT473732 |
| SP | prepared from *E. japonica*-1 | – | – | MT473731 |
| PPY-1 | Kangmei Pharmaceutical Co., Ltd, Guangdong | – | YC20181201 | MT473727 |
| PPY-2 | Dongfang Pharmacy, Guangzhou | – | YC20181202 | MT473728 |
| PPY-3 | Henglu Pharmacy, Guangzhou | – | YC20181203 | MT473730 |

### Sequencing, genome assembly and annotation

Approximately 1 μg genomic DNA was randomly fragmented by Covaris (E210), followed by fragments selection by Agencourt AMPure XP-Medium kit to an average size of 200–400 bp. Selected fragments were end-repaired and 3'adenylated, and the resulting DNA was ligated with adaptors. After the ligation, the products were amplified by PCR and purified using Agencourt AMPure XP-Medium kit. The purified double-stranded PCR products were heat-denatured to single stand and circularized by the splint oligo sequence to generate a single strand circular DNA (ssCir DNA) library after quality control. The ssCir DNA molecule formed a DNA nanoball (DNB), and the final DNB was loaded onto a sequencing chip and were sequenced using the BGISEQ-500 platform. Finally, the pair-end (PE) 124–150 bp reads were obtained by combinatorial Probe-Anchor Synthesis (cPAS).

Low-quality reads, adapter contamination, and duplicated reads were removed from the PE sequence data generated from the BGI platform using SOAPnuke software v2.1.5 [57] to produce the "clean data", which were filtered using Bowtie2 [58] and then assembled using SPAdes v3.14.0 [59] in GetOrganelle v1.7.0 [60]. In cases of failure of ribosomal DNA assembly, we amplified and sequenced the ribosomal DNA to obtain the ITS and ITS2 sequences. To improve genome assembly, we also conducted reference-based genome assembly using the cpDNA sequences available in GenBank (Table 1). The contigs obtained from the GetOrganelle assemblies were aligned to the reference genome, and the aligned contigs were assembled to each cpDNA in Geneious v2020.0.4 [61].

The assembled cpDNAs were annotated using GeSeq (Annotation of Organellar Genomes) (https://chlorobox.mpimp-golm.mpg.de/geseq.html) [62] and Plastid Genome Annotator (PGA) [63] software, followed by manual adjustments of the start and stop codons and the exon and intron boundaries via Geneious. The ribosomal DNA was annotated using Geneious. All the tRNA genes were confirmed using the online tRNAscan-SE v2.0.7 [64, 65] and ARAGORN v1.2.38 [66]. The OGDRAM (http://ogdraw.mpimp-golm.mpg.de/) [67] software was used to draw the circular cpDNA maps. The annotated cpDNAs and the ribosomal DNA sequence were submitted to GenBank (http://www.ncbi.nlm.nih.gov/) to obtain the accession number (Table 2). The IR and SSC boundary regions of *E. japonica* species were compared and examined with other cpDNAs.

### Genome structure and comparative analysis

CpDNA characteristics (e. g. structure and genes; codon usage, SSRs and long repeat sequences) were compared among the species concerned for species identification. To determine whether the chloroplast genome sequences of PPY and SP obtained herein were complete, we also compared cpDNA characteristics between PPY/SP and fresh samples. The codon usage and the relative synonymous codon usage values (RSCU) of cpDNAs exons in the consensus protein-coding genes of each species were obtained using CondoW v1.4.2 [68]. The MISA software v2.1 [69] was used to predict the simple repeats (SSR) in cpDNA using the following parameter setting: mononucleotide repeat number > 10, dinucleotide repeat number > 5, trinucleotide repeat number > 4, tetranucleotide, pentanucleotide and hexanucleotide repeat number > 3; the minimum distance between two SSRs was set as 100 bp. If the distance between two SSRs was less than 100 bp, the two SSRs were regarded as one composite microsatellite. The Tandem Repeats Finder was used to predict the tandem repeats with parameters of 2 for the matching weight, 5 for the penalty on the mismatching and the indel, the minimum alignment score to report repeat was set to 50, and 500 for the maximum period size to report [70]. Repeat sequences were predicted by the website REPuter [71]. The minimum repeat size was set to 30 bp, and the sequence identity with Hamming distance was 3. The cpDNA of *E. japonica* was used as the reference sequence, and the sequence similarity of cpDNA was analyzed by Shuffle-LAGAN mode of mVISTA [72].

### Phylogenetic analysis and tree-based identification

The identification capability of cpDNA and the universal barcode regions were compared by constructing a maximum likelihood (ML) tree based on complete cpDNA, ITS and ITS2. Additional nine ITS sequences, two ITS2 sequences and eight cpDNA sequences were also downloaded from GenBank (Additional file 1: Table S4) to enrich the data set. The cpDNAs, ITS, and ITS2 sequences of all species in this study and the published genomes from GenBank were aligned using MAFFT v7.037 [73] and adjusted manually with MEGA6 software as needed [74]. The cpDNA sequences downloaded from GenBank were listed in Table 1. The best-fit substitution models for these cpDNA sequences were inferred by ModelFinder [75] integrated into PhyloSuite [76] based on the Akaike Information Criterion (AIC). Phylogenetic trees were constructed by ML using RAxML (v8.2.4) with the GTR + F + G4 model [75] and 1000 bootstrap replicates. The genetic distance between the species in this study and the reference sequences mentioned above was calculated based on the Kimura 2-parameter distance (K2P) model [77].

### Abbreviations

cpDNA: Chloroplast DNA; ITS: Internal transcribed spacer; IR: Inverted repeat; LSC: Large single-copy; ML: Maximum likelihood; SSC: Small single-copy.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12896-021-00728-0.

**Additional file 1: Fig. S1.** Phylogenetic tree constructed using ML tree based on 20 ITS2 sequences. The number above the branches are bootstrap support values. **Table S1.** Interspecific (below diagonal) and intraspecific (diagonal) genetic distance of cp genomes of six species. **Table S2.** Interspecific (below diagonal) and intraspecific (diagonal) genetic distance of ITS of six species. **Table S3.** Interspecific (below diagonal) and intraspecific (diagonal) genetic distance of ITS2 of six species. **Table S4.** Additional ITS/ITS2 and cpDNA sequences downloaded from the GenBank to construct ML tree.

### Acknowledgements

The authors would like to express their sincere appreciation to Dr. Zhang Yanjun of Wuhan Botanical Garden for gifting the materials of *E. cavaleriei*.

### Authors' contributions

ZC conceived the study and provided the funding. FL, ET and CL collected samples and conducted the experiment. XX performed the genome assembly and analysis on the data. FL conducted the experiment and wrote the manuscript. RH, and ZC gave the comments on the manuscript. All authors have read and approved the final manuscript.

### Funding

This work was partly supported by the National Natural Science Foundation of China (NSFC Grant No. 81373905), Guangdong Provincial Special Fund for Modern Agriculture Industry Technology Innovation Teams, China (2019KJ148) and Natural Science Foundation of Guangdong Province (No. 2020A1515010627).

### Availability of data and materials

The complete chloroplast genomes of 11 samples were submitted to the NCBI database (https://www.ncbi.nlm.nih.gov/). All other data and material 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 that they have no competing interests.

### Author details

[1]Department of Pharmacy, Zhujiang Hospital, Southern Medical University, Guangzhou 510282, China. [2]Faculty of Medicinal Plants and Pharmacognosy, School of Traditional Chinese Medicine, Southern Medical University, Guangzhou 510515, China. [3]Guangdong Provincial Key Laboratory of Chinese Medicine Pharmaceutics, Guangzhou 510515, China.

Received: 5 March 2021 Accepted: 6 December 2021
Published online: 11 December 2021

### References

1. Jiang P, Shi FX, Li MR, Liu B, Wen J, Xiao HX, et al. Positive selection driving cytoplasmic genome evolution of the medicinally important ginseng plant genus *Panax*. Front Plant Sci. 2018;9:359. https://doi.org/10.3389/fpls.2018.00359.
2. Takeshi T, Marouane B, Takuya I, Kazusato O, Kimiko I, Takayuki O, et al. Optimized method of extracting rice chloroplast DNA for high-quality plastome resequencing and *de novo* assembly. Front Plant Sci. 2018;9:266. https://doi.org/10.3389/fpls.2018.00266.
3. Wang WW, Schalamun M, Morales-Suarez A, Kainer D, Schwessinger B, Lanfear R. Assembly of chloroplast genomes with long- and short-read data: a comparison of approaches using *Eucalyptus pauciflora* as a test case. BMC Genomics. 2018;19:977. https://doi.org/10.1186/s12864-018-5348-8.
4. Xing SC, Liu CJ. Progress in chloroplast genome analysis. Prog Biochem Biophys. 2008;35:21–8.

5. 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. https://doi.org/10.1007/s11103-011-9762-4.
6. Ruhlman TA, Jansen RK. The plastid genomes of flowering plants. Methods Mol Biol. 2014;1132:3–38. https://doi.org/10.1007/978-1-62703-995-6_1.
7. Jiang WJ, Guo MY, Pang XH. Application of chloroplast genome in identification and phylogenetic analysis of medicinal plants. World Chin Med. 2020;15(702–708):716. https://doi.org/10.3969/j.issn.1673-7202.2020.05.008.
8. Clegg MT, Gaut BS, Learn GH, Morton BR. Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci. 1994;91:6795–801. https://doi.org/10.1073/pnas.91.15.6795.
9. 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;10:3262. https://doi.org/10.3390/ijms19103262.
10. Asaf S, Khan AL, Khan A, Al-Harrasi A. Unraveling the chloroplast genomes of two prosopis species to identify its genomic information, comparative analyses and phylogenetic relationship. Int J Mol Sci. 2020;21:2380. https://doi.org/10.3390/ijms21093280.
11. Kyalo CM, Li ZZ, Mbandi M, Malombe I, Hu GW, Wang QF. The first glimpse of *Streptocarpus ionanthus* (Gesneriaceae) phylogenomics: analysis of five subspecies' chloroplast genomes. Plants. 2020;9:456. https://doi.org/10.3390/plants9040456.
12. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7:84. https://doi.org/10.1186/1741-7007-7-84.
13. Zhang YJ, Li DZ. Advances in phylogenomics based on complete chloroplast genomes. Plant Divers Resour. 2011;33:365–75. https://doi.org/10.3724/SP.J.1143.2011.10202.
14. Sun YX, Moore MJ, Zhang SJ, Soltis PS, Soltis DE, Zhao T, et al. Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. Mol Phylogenet Evol. 2016;96:93–101. https://doi.org/10.1016/j.ympev.2015.12.006.
15. Yan MH, Fritsch PW, Moore MJ, Feng T, Meng A, Yang J, et al. Plastid phylogenomics resolves infrafamilial relationships of the Styracaceae and sheds light on the backbone relationships of the Ericales. Mol Phylogenet Evol. 2018;121:198–211. https://doi.org/10.1016/j.ympev.2018.01.004.
16. Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot. 2012;99:349–64. https://doi.org/10.3732/ajb.1100335.
17. Dodsworth S. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 2015;20:525–7. https://doi.org/10.1016/j.tplants.2015.06.012.
18. Moore MJ, Bell CD, Soltis PS, Soltis DE. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA. 2007;104:19363–8. https://doi.org/10.1073/pnas.0708072104.
19. Gao L, Su YJ, Wang T. Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. J Syst Evol. 2010;48:77–93. https://doi.org/10.1111/j.1759-6831.2010.00071.x.
20. Wang A, Wu H, Zhu X, Lin J. Species Identification of *Conyza bonariensis* assisted by chloroplast genome sequencing. Front Genet. 2018;9:374. https://doi.org/10.3389/fgene.2018.00374.
21. Abdulah, Mehmood F, Shahzadi I, Waseem S, Mirza B, Ahmed I, et al. Chloroplast genome of *hibiscus rosa-sinensis* (Malvaceae): comparative analyses and identification of mutational hotspots. Genomics. 2020;112:581–91. https://doi.org/10.1016/j.ygeno.2019.04.010.
22. Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N, Cordeiro GM, et al. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J. 2010;9:328–33. https://doi.org/10.1111/j.1467-7652.2010.00558.x.
23. Li XW, Hu ZG, Lin XH, Li Q, Gao HH, Luo GA, 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.
24. Li XW, Gao HH, Wang YT, Song JY, Henry R, Wu HZ, et al. Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species. Sci China Life Sci. 2013;56:189–98. https://doi.org/10.1007/s11427-012-4430-8.
25. Dan J, Zhao ZY, Zhang T, Zhong WH, Liu CS, Yuan QJ, et al. The chloroplast genome sequence of *Scutellaria baicalensis* provides insight into intraspecific and interspecific chloroplast genome diversity in *Scutellaria*. Genes. 2017;8:227. https://doi.org/10.3390/genes8090227.
26. Chen Q, Wu XB, Zhang DQ. Phylogenetic analysis of *Fritillaria cirrhosa* D. Don and its closely related species based on complete chloroplast genomes. Peer J. 2019;7:e7480. https://doi.org/10.7717/peerj.7480.
27. Chen XL, Zhou JG, Cui YX, Wang Y, Duan BZ, Yao H. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. Front Pharmacol. 2018;9:695. https://doi.org/10.3389/fphar.2018.00695.
28. Dobrogojski J, Adamiec M, Luciński R. The chloroplast genome: a review. Acta Physiol Plant. 2020;42:98. https://doi.org/10.1007/s11738-020-03089-x.
29. Adams RP, Sharma LN. DNA from herbarium specimens: I. Correlation of DNA size with specimen age. Phytologia. 2010;92:346–53.
30. Weiβ CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R Soc Open Sci. 2016;3:160239. https://doi.org/10.1098/rsos.160239.
31. Chinese Pharmacopoeia Commission. Pharmacopoeia of the Peoples Republic of China. China Medical Science Press; 2020. p. 213–4.
32. The Wall Street Journal. Herbal supplement has some New Yorkers talking, instead of coughing. https://www.wsj.com/articles/herbal-supplement-has-some-new-yorkers-talking-instead-of-coughing-1519316304. Accessed 20 Sept 2020.
33. Lin YL, Lin WJ, Lin LQ. Research status and development prospect of loquat leaves. J Chin Med Mater. 2006;29:1111–4. https://doi.org/10.3321/j.issn:1001-4454.2006.10.046.
34. Wei JF, Huang CX. Identification of loquat leaf and its counterfeit. Chin Med Bull. 1986;11:72.
35. Zhang ZK, Li GF, Yang XH, Lin SQ. Taxonomic studies using multivariate analysis of *Eriobotrya* based on morphological traits. Phytotaxa. 2017;302:122.
36. Gisbert AD, Romero C, Martínez-Calvo J, Leida C, Llácer G, et al. Genetic diversity evaluation of a loquat (*Eriobotrya japonica* (thunb) lindl) germplasm collection by SSRs and s-allele fragments. Euphytica. 2009;168:121–34.
37. Ren WG, Liu DW, Lin SS, Li WT, Huang LF. Determination of anti-EGFR chemical constituents from Eriobotryae Folium and its active parts by UPLC-Q-TOF /MS. Chin J New Drugs. 2013;22:2012–5.
38. Bakker FT. Herbarium genomics: skimming and plastomics from archival specimens. Webbia. 2017;72:35–45. https://doi.org/10.1080/00837792.2017.1313383.
39. Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, Sara VDK, et al. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biol J Linn Soc. 2016;117:33–43. https://doi.org/10.1111/bij.12642.
40. Zeng CX, Hollingsworth PM, Yang J, He ZS, Zhang ZR, Li DZ, et al. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods. 2018;14:43. https://doi.org/10.1186/s13007-018-0300-0.
41. Alsos IG, Lavergne S, Merkel MKF, Boleda M, Coissac E. The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried material. Plants. 2020;9:432. https://doi.org/10.3390/plants9040432.
42. Chen S, Yao H, Han JP, Liu C, Song JY, Zhu YJ, et al. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE. 2010;5: e8613. https://doi.org/10.1371/journal.pone.0008613.
43. Chen SL, Yao H, Han JP, Xin TY, Pang XH, Shi LC, et al. Principles for molecular identification of traditional Chinese materia medica using DNA barcoding. China J Chin Mater Med. 2013;38:141–8. https://doi.org/10.4268/cjcmm20130201.
44. Hillis DM, Dixon MT. Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol. 1991;66:411–53. https://doi.org/10.1086/417338.

45. Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann Mo Bot Gard. 1995;82:247–77. https://doi.org/10.2307/2399880.
46. Yuan QJ, Zhang B, Jiang D, Zhang WJ, Lin TY, Wang NH, et al. Identification of species and materia medica within *Angelica* L. (Umbelliferae) based on phylogeny inferred from DNA barcodes. Mol Ecol Resour. 2015;15:358–71. https://doi.org/10.1111/1755-0998.12296.
47. Cheng T, Xu C, Lei L, Li CH, Zhang Y, Zhou SL. Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. Mol Ecol Resour. 2016;16:138–49. https://doi.org/10.1111/1755-0998.12438.
48. Fu XM. The relationship of *Eriobotrya* Lindl in Guangdong Province. Doctoral dissertation. 2006
49. Yang XH, Li P, Liu CM, Lin SQ. Genetic diversity in *Eriobotrya* genus and its closely related plant species using RAPD markers. J Fruit Sci. 2009;26:55–9. https://doi.org/10.13925/j.cnki.gskk.2009.01.011.
50. Zhang Y, Qin LH, Wang HK, Chen XP, Wang SH. Identification of S genotypes in loquat (*Eriobotrya japonica* Lindl.) based on allele specific PCR. Sci Hortic. 2017;225:736–42.
51. Sass C, Little DP, Stevenson DW, Specht CD. DNA barcoding in the Cycadales: testing the potential of proposed barcoding markers for species identification of Cycads. PLoS ONE. 2007;2: e1154. https://doi.org/10.1371/journal.pone.0001154.
52. Chase MW, Fay MF. Barcoding of plants and fungi. Science. 2009;325:682–3. https://doi.org/10.1126/science.1176906.
53. Tonti-Filippini J, Nevill PG, Dixon K, Small I. What can we do with 1000 plastid genomes? Plant J. 2017;90:808–18. https://doi.org/10.1111/tpj.13491.
54. Hollingsworth P, Li DZ, Bank M, Twyford A. Telling plant species apart with DNA: from barcodes to genomes. Philos Trans R Soc B. 2016;371:1–9. https://doi.org/10.1098/rstb.2015.0338.
55. Yang JB, Li DZ, Li HT. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. Mol Ecol Res. 2014;14:1024–31. https://doi.org/10.1111/1755-0998.12251.
56. Zhong Y, Du QZ, Li HF, Meng XX, Yuan WM, Wang HK, et al. The nuclear DNA isolation from the leaves of loquat and analysis of its SSR. J Anhui Agri Sci. 2010;38(9419–9422):9434.
57. Chen YX, Chen YS, Shi CM, Huang ZB, Zhang Y, Li SK, et al. SOAPnuke: a mapreduce acceleration supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience. 2018;7:1–6. https://doi.org/10.1093/gigascience/gix120.
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
59. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021.
60. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. Genome Biol. 2019;21:241. https://doi.org/10.1101/256479.
61. Kearse M, Moir R, Wilson A, Stones-Havas 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;28:1647–9. https://doi.org/10.1093/bioinformatics/bts199.
62. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq—versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6–11. https://doi.org/10.1093/nar/gkx391.
63. Qu XJ, Moore MJ, Li DZ, Yi TS. PGA: a software package for rapid, accurate, and fexible batch annotation of plastomes. Plant Methods. 2019;15:50. https://doi.org/10.1186/s13007-019-0435-7.
64. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA Genes in genomic sequence. Nucleic Acids Res. 1997;25:0955–64. https://doi.org/10.1093/nar/25.5.0955.
65. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:686–9. https://doi.org/10.1093/nar/gki366.
66. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32:11–6. https://doi.org/10.1093/nar/gkh152.
67. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004;20:3252–5. https://doi.org/10.1093/bioinformatics/bth352.
68. Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15:1281–95. https://doi.org/10.1093/nar/15.3.1281.
69. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 2003;106:411–22. https://doi.org/10.1007/s00122-002-1031-0.
70. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80. https://doi.org/10.1093/nar/27.2.573.
71. 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. https://doi.org/10.1093/nar/29.22.4633.
72. Frazer KA, Lior P, Alexander P, Rubin EM, Inna D. Vista: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273–9. https://doi.org/10.1093/nar/gkh458.
73. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.
74. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:2725–9. https://doi.org/10.1093/molbev/mst197.
75. Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler AV, Jermiin L. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9. https://doi.org/10.1038/nmeth.4285.
76. Zhang D, Gao F, Jakovli I, Zou H, Wang GT. Phylosuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20:348–55. https://doi.org/10.1111/1755-0998.13096.
77. Kimura M. A simple method of estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20. https://doi.org/10.1007/BF01731581.

## Publisher's Note

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