# Comparative Chloroplast Genome Analysis of *Impatiens* Species (Balsaminaceae) in the Karst Area of China: Insights Into Genome Evolution and Phylogenomic Implications.

**Chao Luo**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**WuLue Huang**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Huayu Sun**
Department of Landscape Architecture and Plant Science, University of Connecticut, Storrs, CT, 06269
**Huseyin Yer**
Department of Landscape Architecture and Plant Science, University of Connecticut, Storrs, CT, 06269
**Xinyi Li**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Yang Li**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Bo Yan**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Qiong Wang**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Yonghui Wen**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Meijuan Huang**
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming
**Haiquan Huang** ( haiquanl@163.com )
College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming

**Research Article**

**Keywords:** Impatiens, Balsaminaceae, chloroplast genome, comparative analysis, phylogenetic relationship

**Posted Date:** January 4th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-132878/v1



**Version of Record:** A version of this preprint was published at BMC Genomics on July 24th, 2021. See the published version at https://doi.org/10.1186/s12864-021-07807-8.

# Abstract

**Background:** *Impatiens,* a controversial and complex genus, which belongs to the family Balsaminaceae with approximately 1000 species. The genus is well known for economical, medicinal, ornamental, and horticultural values. However, Due to the morphological features and insufficient genomic resources, their analyses of germplasm identification and molecular phylogeny are very limited.

**Results:** We have sequenced the chloroplast genomes of six different species (*I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis,* and *I. stenosepala*) in the karst area of China and compared them with previously published species and the monospecific sister *Hydrocera triflora* belonging to Balsaminaceae family. We comtrasted the genome features, repeat sequences, sequence divergence, and constructed the phylogenetic relationships. The 12 complete chloroplast genomes of the Balsaminaceae species ranged in size from 151,538 bp (*I. fanjingshanica*) to 154,189bp (*H. triflora* ) encoded 114 total distinct genes except for *I. glandulifera* and *H. triflora*, including 81 protein-coding, 29 transfer RNA genes(tRNA), and 4 ribosomal RNA genes (rRNA). Moreover, the characteristics of the long repeats sequences and simple sequence repeats (SSRs) were found. Divergent hotspots regions *psbK-psbI,trnT-GGU-psbD, rpl36-rps8, rpoB-trnC-GCA, trnK-UUU-rps16, trnQ-UUG, trnP-UGG-psaJ, trnT-UGU-trnL-UAA, and ycf4-cemA* were identified in the 12 Balsaminaceae chloroplast genomes, which could be suitable for species identification and phylogenetic studies. Additionally, phylogenetic relationships based on Maximum likelihood (ML) and Bayesian inference (BI) among whole chloroplast genomes showed that *H. triflora* was the basal group in Balsaminaceae and *I. guizhouensis* was the basal group in *impatiens* species. Besides this, cultivated species(*I. balsamina, I. hawkeri,* and *I. walleriana*) were clustered together.

**Conclusion:** Our study provides detailed information about nucleotide diversity hotspots and characterized types of repeats, which can be used for developing molecular markers applicable in Balsaminaceae species. And also we reconstructed and analyzed the relationships of some impatiens species and discovered their status based on the complete chloroplast genomes. The current study might provide valuable significant genomic information for the systematics and evolution in the Balsaminaceae.


# Background

The nuclear, chloroplast(cp), and mitochondria genomes are the three major organelles that contain their genomes[1]. Typically, cp genomes in angiosperms comprise a quadripartite circular double-helix structure of highly conserved sizes, structures, and gene sequences ranging from 115 kb to 165 kb[2]. The common feature of the complete chloroplast genome is a typical tetrad structure, which consists of a pair of inverted repeats(IRs), which are divided by large and small single copy(LSC, SSC). Generally, cp genomes contain 110-113 genes, The genes are separated by three categories according to their functions[3]. The first is related to the expression of chloroplast genes, such as tRNA, rRNA, and the three subunits that encode RNA polymerase synthesis. The following is photosynthesis, and the last is composed of other biosynthetic genes and some unknown functions genes, such as *matK* and *ycf1*[4]. Comparison to the nuclear and mitochondrial genomes, the whole chloroplast genome has its self-replication mechanism, relatively independent evolution and small genome size, slow evolving nature and mutation rate, and unique maternal inheritance[5], it is feasible for providing pieces of information for the evolution, DNA barcoding and also reconstructing of plant phylogeny and taxonomy between families and genera from the perspective of population genetics to investigate deep comparisons of angiosperm, gymnosperm, and fern families[6].

Furthermore, mutations, rearrangements, duplications, and losses of genes could be observed in the cp genomes of angiosperm lineages[7]. The structural changes in genomes can be used to study the taxonomic significance and phylogenetic relationship[8]. And also serve as supplied information for developing genomic markers for the complexes taxonomically challenging species[9]. Complete chloroplast contains all genes for the reconstruction of evolutionary history and can provide more valuable information and higher quality for the evolutionary and phylogenetic analyses[10]. And also can reduce the sampling error inherent in analyses of one or a few genes that may indicate crucial evolutionary events. Entire chloroplasts or

protein-coding genes as a powerful and simple strategy can provide valuable information and stronger phylogeny evidence for the identification, classification, and phylogenetic reconstruction among species and families[11].

*Impatiens* species, belonging to the Balsaminaceae, are the controversial and complex classification flowering genus which have been widely used as medicinal, ornamental, and horticultural plants in the regions of North America, Europe, and China[12]. The family of Balsaminacea consists of only two genera of *Impatiens* and the monospecific sister *Hydrocera triflora*(GenBank KF986530.1), with strong similarity in morphology and molecular biology datasets[13]. *Impatiens* are about 1000 species, distributing from tropics to subtropics and extending to sea level to an altitude of 4,000 meters[14], which Tropical Africa, Madagascar, Sri Lanka, Himalayas, and south-east Asia are the five biodiversity hotspots for the endemic *Impatiens*[15].

The center of origin and diversification of Balsaminaceae in China[16]. Especially in the karst area of Southwest China, approximately 250 wild impatiens species were described from Guizhou, Yunan, Sichuan, and Tibet, many of which are used as supplements for medicinal or health purposes. In ancient China, *Impatiens* was called 'zhijiahua' which can be used to crush into mashes and directly applied on the nails[17][18]. The pharmaceutical and chemical products can be used as annual herbs for the medical treatment of rheumatism, beriberi, bruises, pain, warting, snakebite, fingernail inflammation, and onychomycosis[19][20]. Additionally, high levels of metals such as copper, zinc, chromium, and nickel can be accumulated by the *Impatiens* species based on the previous research [21].

Due to the flowering diversely and morphologically variable, the phylogenetic relationships of impatiens plants are still unknown[22]. Impatiens are characterized by zygomorphic flowers with enormous diversity and high levels of convergent evolution variability in corolla color and morphology, the flowers are extremely fragile, and most are coalesced and folded in dried specimens, making it difficult to separate and reconstruct different parts laborious[23][24]. Meanwhile, due to the semi-succulent stems and many fleshy leaves, it is difficult to provide well-dried herbarium plant specimens[25]. Early research on *Impatiens* was primarily focused on a specific geographical area and providing purely descriptive traditional taxonomy processing[26]. To date, the only global infrageneric molecular classification for *Impatiens* was done based on several plastids (such as coding gene *matK, rbcL, trnK,* and intergenic region *atpB-rbcL* and *trnL-trnF*) [27][28]. And also, the nuclear ribosomal ITS and the inter-simple sequence repeat (ISSR) markers were used to identify the genetic diversity of populations and help to understand the phylogenetic and evolutionary relationship between species of *Impatiens*[29]. However, all the published data contained only a few samples from obvious regional characteristics, and also some species with diversified morphology have taxonomic controversy due to unresolved phylogenetic relationships and require further studies and clarification[30]. Therefore, using whole chloroplast genomes as the sequence is remarkably to increase and improve the additional evidence for identifying species and reconstructing the phylogeny relationships[31].

12 new complete chloroplast genomes of Impatiens were assembled, which including six firstly sequenced chloroplast genomes (*I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis,* and *I. stenosepala*) in the karst area of China and combined with previously published Balsaminaceae complete chloroplast genomes by using Illumina sequencing technology[32]. The present investigation is a novel attempt to reveal and identify the phylogenetic analysis of the taxonomic position of *Impatiens* based on the whole chloroplast genome. The aims of this study are: (i) to conduct a comprehensive analysis of the pomegranate chloroplast genome, including basic genome structure information, codon usage, repetitive structure characteristics, and IR region expansion, (ii) to identify hotspot regions, microsatellite types, and comparative genomic divergence; and (iii) to reconstruct and analyze the relationships of the impatiens species and find the status of Impatiens based on the complete chloroplast genomes.

# Results

### General features of *Impatiens*

As a result, the genomic libraries have a total of 28.6 GB. The 12 complete Balsaminaceae species cp genomes ranged in size from 151,538 bp (*I. fanjingshanica*) to 154,189bp (*H. triflora* ) (Table 1). The newly sequenced impatiens cp genomes maps

were provided in Fig. 1 and Supplementary Fig. S1-S6 *(I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis, and I. stenosepala)*. Similar to other typical chloroplast genomes of angiosperms, the common feature of the complete cp genomes consisted of four conjoined regions forming a circular molecular structure. The pair of inverted repeats are separated by a large and small single copy(LSC, SSC). In the family Balsaminaceae, The LSC constituted for 54.47-55.04% of the total chloroplast genome size, ranging from 82,542 bp (*I. fanjingshanica*) to 84,865 bp (*H. triflora*); The SSC constituted for11.37-11.73% of the total chloroplast genome size, ranging from 17,309 bp (*I. linearisepala*) to 18,080 bp (*H. triflora*); The IR constituted for 16.62-16.98% of the total chloroplast genome size, ranging from 25,622 bp *(H. triflora*) to 25,726 bp (*I. fanjingshanica*). In the newly sequenced genus Impatiens, The LSC constituted for 54.47-54.86% of the total chloroplast genome size, ranging from 82,542 bp (*I. fanjingshanica*) to 83,508 bp (*I. linearisepala*); The SSC constituted for 53.58-58.27% of the total chloroplast genome size, ranging from 17,309 bp (*I. linearisepala*) to 17,547 bp (*I. fanjingshanica*); The IR constituted for 16.83-16.98% of the total chloroplast genome size, ranging from 25,720 bp (*I.stenosepala*) to 25,726 bp (*I. fanjingshanica*).

Like other typical angiosperms, the chloroplast genomes of the Balsaminaceae species encoded 114 total distinct genes except for *I. glandulifera* and *H. triflora*, including 81 protein-coding, 29 transfer RNA genes(tRNA), and 4 ribosomal RNA genes (rRNA) (Supplementary Table S2). But the trnG-UCC gene was annotated as a pseudogene in *H. triflora* comparing to the other *Impatiens* species in a total number of 115 genes. The gene names of the *ycf15* and *trnfM-CAU* gene are interchanged due to incorrect annotation in *I. glandulifera*. These genes were classified into three groups based on the functions: (1) photosynthesis-related genes encoding (Rubisco, ATP synthase, Photosystem I, Cytochrome b/f complex, Photosystem II, Cytochrome c synthesis, and NADPH dehydrogenase). (2) Transcription and RNA genes including 4 transcription genes (*rpoA, rpoB, rpoC1*, and rpoC2*), 20 ribosomal proteins, 4 ribosomal RNA (*rrn4.5, rrn5, rrn16, and rrn23*), and 30 transfer RNA. (3) Other genes including four genes(*matK*,*cemA*,*accD*, and*clpP*) with known function and conserved reading frames (*ycf1, ycf2,* and *ycf15*) encoding proteins (Table 2 and Supplementary Table S1).

16 unique genes were annotated in *Impatiens* species, whereas introns are missing in one of these genes in *I.piufanensis* and *H. triflora*, namely the *rps16* gene and *trnG-GCC* tRNA gene. Among the 16 genes, two genes (*ycf3* and *clpP*) contained two introns and 14 genes contained single intron. Moreover, there were 11 genes(*clpP, ycf3, trnv-UAC, rps12, trnK-UUU,rpoC1, petB, trnL-UAA, atpF, trnG-GCC,*and *rps16*)in the LSC regions, 4 genes(*tRNA-GAU, trnA-UGC, ndhB,* and *rpl2*) in the IR regions and only one gene(*ndhA*) in the SSC regions. The longest intron is the *trnK-UUU*, which is ranging from 2,488 bp (*I. loulanensis*) to 2,548 bp (*I. guizhouensis)*; and the exon of *rpoC1* is the longest. The *rps12* is a trans-splicing gene in which is divided into 5′ -*rps12* in the LSC region and 3′-*rps12* in the IR region (Table 2 and Supplementary Table S3).

**Table 1** Complete chloroplast genomes of six firstly sequenced Balsaminaceae species

| | *I. chlorosepala* | *I. fanjingshanica* | *I. guizhouensis* | *I. linearisepala* | *I. loulanensis* | *I. stenosepala* |
|---|---|---|---|---|---|---|
| reference | this study | this study | this study | this study | this study | this study |
| Family | Balsaminaceae | Balsaminaceae | Balsaminaceae | Balsaminaceae | Balsaminaceae | Balsaminaceae |
| Genus | Impatiens | Impatiens | Impatiens | Impatiens | Impatiens | Impatiens |
| Total length(bp) | 152,763 | 151,538 | 152,774 | 152,212 | 152,472 | 152,802 |
| GC(%) | 36.7 | 36.9 | 37 | 37 | 36.7 | 36.9 |
| LSC length(bp) | 83,740 | 82,542 | 83,572 | 83,508 | 83,460 | 83,626 |
| GC(%) | 34.3 | 34.6 | 34.8 | 34.8 | 34.4 | 34.5 |
| SSC length(bp) | 17,477 | 17,547 | 17,662 | 17,309 | 17,541 | 17,739 |
| GC(%) | 29.5 | 29.4 | 29.9 | 30 | 29.6 | 29.8 |
| IR length(bp) | 25,773 | 25,726 | 25,772 | 25,699 | 25,737 | 25,720 |
| GC(%) | 43.1 | 43.1 | 43 | 43 | 43 | 43.2 |
| CDS length(bp) | 79,562 | 79,689 | 79,941 | 79,533 | 79,650 | 79,581 |
| GC(%) | 37.2 | 37.2 | 37.4 | 37.3 | 37.1 | 37.2 |
| rRNA Length (bp) | 9,048 | 9,048 | 9,046 | 9,048 | 9,048 | 9,048 |
| GC(% | 55.1 | 55.1 | 55.1 | 55.2 | 55.1 | 55 |
| tRNA Length (bp) | 2,876 | 2,872 | 2,872 | 2,872 | 2,872 | 2,884 |
| GC(%) | 52.4 | 52.6 | 52.7 | 52.5 | 52.6 | 52.6 |
| Total Genes | 114 | 114 | 114 | 114 | 114 | 114 |
| CDS | 81 | 81 | 81 | 81 | 81 | 81 |
| tRNA | 29 | 29 | 29 | 29 | 29 | 29 |
| rRNA | 4 | 4 | 4 | 4 | 4 | 4 |

### Differences Genome Size

Among the 12 Balsaminaceae species, the shortest genome was *I. fanjingshanica* (151,538 bp), and the longest was *H. triflora* (154,189 bp). In the 6 new sequenced species, *I. stenosepala* was the longest genome length (152,802 bp) while that of the shortest was *I. fanjingshanica* (151,538 bp). Except for *I. stenosepala* and *I. fanjingshanica*, the sizes of *Impatiens* species were between 152,212 bp and 152,774 bp (Table 1). Except for *I. fanjingshanica*, lengths of other Balsaminaceae species were longer than 152,000 bp (Supplementary Table S1). In the 12 Balsaminaceae species, The Length of Protein Coding Genes constituted ranged from 79,533 bp (*I. linearisepala*) to 80,952 bp (*H. triflora*), The length of the rRNA constituted 9,048 bp except for *I. guizhouensis*, *I. glandulifera*, and *H. triflora*, which the length is 9,046 bp, 9,050, and 9,046 bp respectively. The

length of the tRNA constituted 2,872 bp except for *I. chlorosepala, I. stenosepala, I. glandulifera,* and *H. triflora*, which the length is 2,876 bp, 2,884 bp, 2,419 bp, and 2,815 bp respectively (Supplementary Table S1). The overall guanine-cytosine (GC) contents of each species were very similar in the whole cp genomes and the same regions of the LSC, SSC, and IRs. The whole GC content in the Balsaminaceae species ranged from 36.7% to 37%, with *I. chlorosepala* and *I. loulanensis* having the lowest and *I. guizhouensis* and *I. linearisepala* having the highest GC content (Table 1). The GC contents in the LSC, SSC, and IR regions are average 34.56%, 29.7%, 43.0%, respectively(Table 1 and Supplementary Table S1).

**Table 2.** The List of genes in the chloroplast genomes of *Impatiens* species

| Function of Genes | Group of Genes | Gene Names |
|---|---|---|
| Photosynthesis-related genes | Rubisco | *rbcL* |
| | Photosystem I | *psaA psaB psaC psaI psaJ* |
| | Assembly and stability of Photosystem I | *ycf3** ycf4* |
| | Photosystem II | *psbA psbB psbC psbD psbE psbF psbH psbI psbJ psbK psbL psbM psbN psbT psbZ* |
| | ATP synthase | *atpA atpB atpE atpF* atpH atpI* |
| | Cytochrome b/f complex | *petA petB* petD petG petL petN* |
| | Cytochrome c synthesis | *ccsA* |
| | NADPH dehydrogenase | *ndhA* ndhB*(2) ndhC ndhD ndhE ndhFndhG ndhH ndhI ndhJ ndhK* |
| Transcription and translation-related genes | Transcription | *rpoA rpoB rpoC1* rpoC2* |
| | Ribosomal proteins | *rpl2*(2) rpl14 rpl16 rpl20 rpl22 rpl23(2) rpl33 rpl36 rps2 rps3 rps4 rps7(2) rps8 rps11 rps12*(2) rps14 rps15 rps16* rps18 rps19(2)* |
| RNA genes | Ribosomal RNA | *rrn4.5 rrn5 rrn16 rrn23* |
| | Transfer RNA | *trnA-UGC(2) trnC-GCA trnD-GUC trnE-UUC trnF-GAA trnfM-CAU trnG-GCC* trnG-UCC trnH-GUG trnI-CAU*(2) trnI-GAU(2) trnK-UUU* trnL-CAA(2) trnL-UAG trnL-UAA* trnM-CAU trnN-GUU(2) trnP-UGG trnQ-UUG trnR-ACG(2) trnR-UCU trnS-GCU trnS-GGA trnS-UGA trnT-GGU trnT-UGU trnV-GAC(2) trnV-UAC* trnW-CCA trnY-GUA* |
| Other genes | RNA processing | *matK* |
| | Carbon metabolism | *cemA* |
| | Fatty acid synthesis | *accD* |
| | Proteolysis | *clpP*** |
| Genes of unknown function | Conserved reading frames | *ycf1 ycf2(2) ycf15(2)* |

(2) indicates the m=number of the repeat unit is 2; *Gene contains one intron; **Gene contains two intron

## Codon Usage

To analyze the genetic information and the relationship between evolution and phylogeny of *Impatiens*, we analyzed the codons in its coding region. All protein-coding genes were encoded by 50,512 (*I. fanjingshanica*) to 51,396 (*H. triflora* ). The termination codons were considered by the UGA, UAG, and UAA. For these Balsaminaceae species(Supplementary Table S4), we found that the most abundant AA was leucine, which is the UUA encoded the highest RSCU (Relative Synonymous Codon Usage) value at approximately 1.92. Tryptophan was the lowest frequency AA in the Balsaminaceae species. All amino acids, except for methionine and tryptophan, have more than one synonymous codon. Among them, leucine, arginine and serine have 6 codons. The results of RSCU in A or T nucleotide frequency at the third codon position was biased toward a higher than G or C nucleotide frequency in the 12 Balsaminaceae species. *I. glandulifera* had 30 codons, which was less frequently used than the expected usage at equilibrium (RSCU < 1). *H. triflora* had 36 codons more frequently used than the rest of *Impatiens* species showed the codon usage bias in 34 codons.

### Repeat Structure Analyst

Of the 12 Balsaminaceae species, 246 long repeats of four types (forward, complement, reverse, and palindromic)using REPuter (Supplementary Table S5). The most common repeat type was forward and palindromic repeats. compliment repeats were only identified in *I. guizhouensis*; reverse repeats were only found in *I. Chlorosepala, I. fanjingshanica, I. linearisepala, I. balsamina,* and *I. hawkeri*, respectively. The range of most copy length was 30-40 bp(Figure 2B). The individual accession with the greatest number of repeats was *I. chlorosepala* with 25, comprising 14 forward, 9 palindromic, and 2 reverse repeats. *I. linearisepala* which had the smallest number of repeats had 5 forward, 7 palindromic, and 3 Reverse repeats(Figure 2A). The greatest numbers of Forward, palindromic, and Reverse repeats were found in *I. chlorosepala* (14), *I.balsamina* (34), and *I. linearisepala* (3), respectively.

### Simple Sequence Repeat Analysis

Simple Sequence Repeat, called microsatellites are widely used as molecular markers and play a significant role in plant identification and classification. Among these SSRs in the 12 Balsaminaceae species, the distribution of 51-109 SSRs ranged in size from 10 to 20 bp. There were 6 kinds of SSRs were discovered (Figure 3A and Supplementary Table S6). Among these SSRs, only *H. triflora* had the hexanucleotide repeats, and *I. loulanensis, I. stenosepala, I.balsamina, I.walleriana,* and *H. triflora* had the Pentanucleotide repeats. The numbers of mononucleotide repeats ranged from 59(*I. linearisepala*) to 82(*I. chlorosepala*), followed by Dinucleotide ranging from 5(*I.hawkeri*) to 13(*I. chlorosepala, I. fanjingshanica, and I. glandulifera*) (Figure 3B-G). Therefore, mononucleotide and Dinucleotide repeats may play a more significant role in genetic variation.

In the six newly species, mononucleotide repeats were more abundant with A/T repeats being the most highly represented repeats, whereas poly C/G repeats were rather rare. poly C/G repeats were only found in *I. chlorosepala, I. fanjingshanica, I. guizhouensis, and I. loulanensis.* Moreover, the number of A mononucleotide repeats ranged from 24(*I. fanjingshanica, and I. linearisepala*) to 37(*I. loulanensis*,), with T mononucleotide repeats ranging from 35 (*I. linearisepala*) to 48 ( *I. fanjingshanica*) (Figure 3B-G).

Among the dinucleotide repeat motifs AT/AT were more abundant. In the newly sequenced species, the SSR analysis showed that *I. chlorosepala* had the highest number of SSRs (109) while *I. linearisepala* had the lowest (74). Trinucleotide motif (ATT, GAA, TAA, TTA, TAT, ATA, and TTG) and tetranucleotide( AAAT, AATA, AATT, ATAA, TAAA, TATT, TTCA, TTTA, GTTT, and TTCT) were identified. However, among these cp genomes, only Pentanucleotide (AAAAG and CAAAA) repeat was found in the *I. loulanensis* and *I. stenosepala* cp genome.

### Comparison of the Genome Structure

The structure and size of the chloroplast genome can change based on the different evolution and genetic backgrounds. The collinear method was used to analyze and compare the chloroplast genomes. Mauve alignment of plastomes shows that the impatiens plastome structure is similar to the dicot Rosa(MK947051.1)(Figure 4A). *But compare with* monocot Triticum aestivum(NC_002762.1) and Oryza sativa(NC_008155.1), the results showed that the monocot and dicot structures derive from intermolecular recombination events(Figure 4A). There are no interspecific and intraspecific rearrangements within six species

revealed that all genes (including ribosomal RNA, tRNA, and protein-coding genes) in the Balsaminaceae is comparatively conserved and presented in the same order(Figure 4B); this also applies optimal collinearity between subgenus *Impatiens*, there is no gene rearrangement. Moreover, Compared with *H. triflora*, the linear relationships with genome structure and gene sequences indicate that there was high chloroplast genome homology.

### Comparative Genomic Divergence and Genome Rearrangement

A comparative cp genomes analysis of the whole regions between *H. triflora* and other impatiens species notably, *I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis,* and *I. stenosepala* was conducted by using the mVISTA software and DnaSP to detect hyper-variable regions and analyze the sequence identity plots between the entire cp genome by sequence identity diagram(Figure 5A). The comparison showed that the number and sequence of genes in the IR regions were relatively conservative and less divergent than the LSC and SSC regions, The IR regions were the most conserved(Figure 5B and 5C).

Among the coding genes, the highly divergent regions such as *matK, psbK, petN, psbM, atpE, rbcL, accD, psaL, rpl16, rpoB, ndhB, ndhF, ycf1,* and *ndhH* (Figure 5A). For the intergenic regions, *atpH-atpI, trnC-trnT, rps3-rps19*, and *ndhG-ndhA* were the most variable. In the LSC region, the *psbK-psbI, atpI*, and *rps4-trnF* genes showed some sequence divergence in *I. piufanensis, I. glandlifera,* and *H. triflora*. The three genes of *ndhF, ycf1*, and *ndhH* were detected in the SSC region. The *rpl32-trnN* showed the highest variation among the hypervariable regions and the *ycf1* gene is the most divergent. In comparison with *H. triflora*, the large copy of the *trnI-trnN* and *trnA-trnL* locus in the cp genomes of *I. fanjingshanica, I. guizhouensis,* and *I. loulanensis* have been deleted.

### Sequence Divergence and Mutational Hotspots

We compared the Pi values in DnaSP 5.1. to determine the divergent hotspots region in 12 Balsaminaceae species, The analysis indicated that the variation in LSC and SSC regions was much higher than that in the IR regions(Figure 6). The highest value of nucleotide diversity (Pi) was *ycf1*(0.17356) and *trnG-GCC*(0.12911). 6 mutational hotspots which exhibited remarkably higher Pi values (>0.06) in LSC and SSC regions were *trnk-UUU-rps16, trnG-GCC, atpH-atpL, rpoB-petN, rps4-ndhJ,* and *accD-psaI,* while in the SSC region with three hotspots(*ndhF, rpl32-ccsA,* and *ycf1*) above 0.06. Similarly, we determined the average pairwise sequence divergence among new sequenced impatiens species. The nucleotide variability (Pi) of these 140 regions ranged from 0. 0% (*rrn16*) to 9.3% (*rps12*). The *rps12* gene demonstrated the highest average sequence divergence (0.93), followed by *rpl32* (0.91), and *rps4-ndhJ*(0.90)(Figure 6 and Supplementary Table S7). By contrast, the Pi values of 6 new sequenced species were higher than those of the 12 Balsaminaceae species. Therefore, these coding regions and non-coding genes may provide strong molecular evidence for resolving future low-level phylogeny and phylogeography in Balsaminaceae.

### Contraction and Expansion of Inverted Repeats (IRs)

The genomic structure, the number and sequence of genes were highly conserved in the 12 Balsaminaceae species. But, the contraction and expansion of IR boundaries changed in structure and the sizes. In the 12 Balsaminaceae species, we found the ycf1 existed in the boundaries of IRa-LSC, the sizes of the IRs of *I. chlorosepala* and *I.balsamina* involucrate was the longest (25,773bp) and that of *H. triflora* was the shortest (25,622 bp). The LSC-IRB junctions were embedded in the rps19 genes. The length of rps19 in the LSC region varied from 0bp to 246bp. However, the overlap between rps19 in the IRb region was varied from 0 bp to 200 bp. The IRB-SSC junction was located or adjacent to gene *ycf1* and *ndhF*; all species except for *I. linearisepala* were all located and adjoined the end of *ycf1* from 0bp to 1256bp, the distances between *ycf1* and IRB-SSC junction in I. linearisepala was 204 bp. The overlap between *ndhF* and *ycf1* gene was detected in *I. guizhouensis, I. linearisepala, and I. Hawkeri*, where *ndhF* expanded into the IRB region for 18bp, 176bp, and 98bp, respectively(Figure 7).

In the other species, the distances between *ndhF* and IRB-SSC junction were varied from 1 bp to 2000 bp. The SSC-IRA junction is located in the pseudogene *ycf1* which covered the IRA and SSC region. the length of pseudogene *ycf1* in the SSC region varied from 4356bp to 4891bp. However, the overlap between pseudogene *ycf1* was varied from 810 bp to 1254 bp in the IRA region. The IRb/SSC and SSC/IRa regions were variable. The *rps19-psbA* coding region intervened in the boundaries of LSC/IRa

regions except for *I.piufanensis, I. glandulifera,* and *H. triflora*, in which there was rps19 gene missing in the junctions of the LSC/IRb regions. However, the length of *rps19* in the LSC region varied from 0 bp to 136 bp, in contrast, the length of rps19 in the IRB region of *I. guizhouensis* and *I. stenosepala* was 31bp, 137bp, respectively. For the 6 new sequenced species(*I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis,* and *I. stenosepala*), the sizes of the IRs of I. chlorosepala was the longest (25,773 bp) and that of I. linearisepala was the shortest (25,699 bp).

### Phylogenetic Analyses Within Balsaminaceae Species

We used the MP and BI phylogenetic trees to explore the phylogenetic positions and evolutionary relationships of Balsaminaceae species based on the complete chloroplast genomes(Supplementary Table S8). These chloroplast genomes from eight families within *impatiens* including four Ebenaceae species, four Styracaceae, five Actinidiaceae species, five Theaceae species, five Primulaceae species, 12 Balsaminaceae species, two Saxifragaceae species, and three Rosaceae species as outgroups. The 12 Balsaminaceae species included three cultivated species(*I. balsamina, I. hawkeri, and I. walleriana*) ,three published plastid genomes(*I. piufanensis, I. glandulifera, and H. triflora*) , and 6 newly sequenced species (*I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis,* and *I. stenosepala*).

The topologies of the two datasets ( ML and BI) generated a high approval rate, and the five selected families (Primulaceae, Actinidiaceae, Theaceae, Ebenaceae, and Styracaceae) were clustered into five monophyletic branches. The Genus Hartia Dunn and the stewartia of the family Ebenaceae were clustered into a clade, While the Theoideae also consisted of the Actinidia and the Rhododendron. The Saxifragaceae families were clustered into a monophyletic branch which is close to the outgroup Rosaceae species. Only three nodes(Primulaceae, Actinidiaceae, and Ebenaceae) with bootstrap values under 90% in the ML tree(Fig. 8A). The remaining nodes had support values 100%. Only two nodes(Primulaceae and Actinidiaceae) with bootstrap values under 90% in the BI tree. The remaining nodes had support values 100% (Fig. 8B).

All Balsaminaceae species formed a monophyletic subclade in ML and BI trees, which was sister to two Saxifragaceae species (*Hydrangea Serrata* and *hydrangea heteromalla*). Moreover, *H. triflora* was initially clustered together with other 11 impatiens species and similar clustering results were also measured in previous studies. The *I. guizhouensis* was the earliest branch and it was sister to the other species in impatiens. Support values in the ML and BI trees showed a sister relationship with cultivated species(*I. balsamina, I. hawkeri, and I. walleriana*) formed a clade with *I. chlorosepala* indicating their close relationship; Besides that, I. fanjingshanica and I.piufanensis formed a close relationship while *I. glandlifera and I. loulanensis* formed a clade with *I. linearisepala and I. stenosepala* with the most similar morphological characteristics were clustered together. As a whole, the BI and ML phylogenetic trees were cleared to reveal the internal relationships among the Balsaminaceae species.

## Discussion

### Chloroplast Genome Structure

12 complete cp genome of Balsaminaceae were compared, which included 108-114 genes composed of 79-81 coding genes, 25-29 tRNAs, and 4 rRNAs. The cp genome of the Balsaminaceae species is a typical tetraploid consisting of two inverted repeat IR regions (large and small single-copy fragments)[33]. The coding genes are divided into three categories according to their functions. The first is related to the expression of chloroplast genes, such as tRNA, rRNA, and the three subunits that encode RNA polymerase synthesis. The following is photosynthesis, and the last is composed of other biosynthetic genes and some unknown functions genes, such as *matK*,*cemA*,*accD*, and *clpP*[34]. The chloroplast genes of the Balsaminaceae are similar in composition. In these chloroplast genomes, 151,538–154,189 bp were generated in length, the overall GC contents were between 34.30 to 34.8%. The *Impatiens* genus with its size varied from 151,538 bp (*I. fanjingshanica*) to 152,802 bp (*I. stenosepala*). There was a 1,336 bp difference in the length between the *Impatiens* species. Compared with H. triflora, the plastomes of impatiens species have reduced in size by approximately1,387-2,651bp. Among them, H. triflora possessed the largest plastome (154,189 bp). Contrarily, in the most reduced plastome (151,538 bp) of *I. fanjingshanica*, the contraction and expansion of IR boundaries have been observed, suggests that the contraction and expansion can be partly responsible for the plastome downsizing in Balsaminaceae species. The potential of the ycf15, trnfM-CAU, and psbN genes had been annotated in

all genomes of Impatiens species, while in H. triflora, they were not excluded in this study. Likewise, the reading frames called; trnG-UCC gene, which had been only annotated in I. glandulifera.

The GC content of *I. chlorosepala* and *I.balsamina* was found to be lower than that of other species(Table 1). The GC content in the IR regions was much higher than in the LSC and SSC regions in all Balsaminaceae species. The four rRNA genes and tRNA have high GC content (52%-55%). Normally, higher GC content indicates more stable genome sequence. This strongly shows that chloroplast genomes have differences in the same family[35]. But the basic structure and content of the genome were roughly similar. And also the chloroplast genomes were highly conserved.

### IR Expansion and Contraction

In most cases, gain or loss genes in cp genomes are usually due to the contraction and expansion of genomes region[36]. The pseudogenes *ycf1* could result in these events. This was apparent in the plastomes of *I. chlorosepala, I. guizhouensis*, and *I. loulanensis* where the IRs were much longer. Interestingly, the cp LSC borders of *I. linearisepala* and *I. guizhouensis* are quite different from other Balsaminaceae species as the *NdhF* gene extended into the IRs and SSC region. And for the IR/LSC junctions of *I. guizhouensis* and *I. stenosepala*, the rps19 gene extended into the IRs and LSC region. The LSC length of *I. fanjingshanica* was the shortest out of the other 12 cp genomes.

### Repetitive-Sequence Analyses and Simple sequence repeats(SSRs) Analyses

Analysis of various cp genomes shows that repetitive sequences are essential for inducing indels and substitutions[37]. It not only plays an important role in the rearrangement and stabilization of the cp genome sequence but also is one of the reasons that affect the copy number differences between similar species and different species[38]. We have identified A total of 246 repeats in Balsaminaceae, falling into four different repetitive categories(Supplementary Table S5). Among all species, the most common type of repetition was palindromic repeats, which occurred 122 times (49.59%), followed by forwarding repeats (114 instances, 46.34%). compliment repeats were only identified in I. guizhouensis. reverse repeats only found in *I. Chlorosepala, I. fanjingshanica*, *I. linearisepala, I. balsamina,* and *I. hawkeri*, respectively.

Simple sequence repeats(SSRs) have been recognized as one of the main sources of molecular markers for having a high polymorphism rate and abundant variation at the species level. Moreover, SSRs are useful for detecting genetic diversity and polymorphisms at the population, intraspecific, and cultivar levels, as well as for distinguishing species[39][40]. A total number of 51-109 SSRs were identified with an overall length ranging from 3 to 10. Additionally, mononucleotide SSRs were detected in all Balsaminaceae with the most frequent, providing ample markers for phylogenetic analysis. The number of SSRs in *H. triflora* were examined lower than *I. chlorosepala*. Poly (A)/(T) SSRs are usually more common than other SSR repeat types, whereas poly C/G repeats were rather rare. We only identified the Hexanucleotide(ATTGGG) in *H. triflora* and poly G in *I. guizhouensis*. We also identified SSR repeats units (AAAGA) unique to *I.balsamina* and *I.walleriana,* and TAAA/TTTA repeat units are unique to *I. chlorosepala* and *H. triflora.* most cpSSRs were observed in non-coding regions, which are generally short mononucleotide tandem repeats and commonly show intraspecific variation in repeat numbers[41]. Because of the slippage of DNA strands, repeat loci, pairwise sequence divergence, and highly divergent regions were detected will be useful tools for investigating levels of genetic diversity and presenting a high mutation rate[42].

### The Utility of Plastomes in Phylogenomics and DNA Barcoding

The divergent hotspots are usually used as evidence for species authentication and provide information about phylogeny[43]. Moreover, the IR regions showed lower sequence divergence than the SSC and LSC regions, and the non-coding regions and the coding regions were less similain the angiosperm cp genomer[44]. Similarly, relatively most divergent genes especially in *trnk-UUU-rps16, trnG-GCC, atpH-atpL, rpoB-petN, rps4-ndhJ, accD-psaI, ndhF, rpl32-ccsA, and ycf1* were detected[45]. Meanwhile, two regions (trnG-GCC and ycf1) showed high levels of variation at both within Balsaminaceae and new sequenced levels (pv >0.8%). For the newly sequenced species, non-coding regions especially possessing high variability as possible molecular

markers such as *psbK-psbI, trnT-GGU-psbD, rpl36-rps8, rpoB-trnC-GCA, trnP-UGG-psaJ, trnT-UGU-trnL-UAA, trnK-UUU-rps16, ycf4-cemA, andtrnQ-UUG* genes.

### Phylogenomic Validation

Balsaminaceae are considered to be a controversial and complex family at both the morphological and molecular levels owing to similar morphology and wide distribution areas[46]. Based on various analyses of the whole chloroplast genome which contain sufficient informative locis for resolving molecular evolution and phylogenetic relationships of the same family and genus, and have been examined to increasingly being used to solve lower taxonomic levels.

The first molecular phylogeny of the genus was published by Fujihashi[47], However, the limited taxon samples and a distant outgroup *Tropaeolum* (Tropaeolaceae) resulted in limited information for the molecular relationships. Nuclear ribosomal internal transcribed spacer (ITS) and *atpB-rbcL* phylogeny study on Balsaminaceae based on sequences for 111 species, provided new phylogenetic insights that *Impatiens* colonized from southwest China to the African continent in three separate proliferation events[48]. Subsequently, based on plastids, combined plastids and nuclear, or combined plastids and pollen data, the *Impatiens* were further analyzed [49]. A new classification of *Impatiens* based on morphological and molecular datasets was divided into two subgenera (*Clavicarpa* and *Impatiens*), seven sections of the subgenus *Impatiens* were further subdivided[50]. Although these results have laid an important foundation for the identification and classification of Balsaminaceae species, However, all the published data contained only a few samples from obvious regional characteristics and the numbers of nuclear/cp gene are relatively less[51], which limits phylogenetic studies and the results were conflicting to provide sufficient information to elucidate the phylogenetic and evolutionary relationships among Balsaminaceae species.

Based on the ML and BI trees (Figure 8), Analysis of the phylogenetic relationships show that the five selected families (Primulaceae, Actinidiaceae, Theaceae, Ebenaceae, and Styracaceae) were clustered into a monophyletic branch, all Balsaminaceae species formed a monophyletic subclade in both trees, which was sister to two Saxifragaceae species (Hydrangea Serrata and hydrangea heteromalla). The relationships within *H. triflora and the* other 11 impatiens species order recovered in this study received high support values (Figure 3), which are highly congruent with previous studies. Cultivate species(I. balsamina, I. hawkeri, and I. walleriana) are included in the same clade, which is far away from the wild species, and we can infer that the evolutionary direction may start the wild *H. triflora to the* cultivated species. And also we can infer that the genes of genetic variations in Balsaminaceae have evolved to adapt to the variety of environments. The structure of *I. guizhouensis* is similar to *H. triflora* than other related impatiens species. Our results suggested that *I. guizhouensis was* the ancestral state in the Balsaminaceae family based on the earliest diverging lineages.

In the morphological classification of *H. triflora*, Leaves alternate, linear-lanceolate, sessile. Sepals 4, unequal length, 5 stamens, 5 ovary, 2 to 3 ovules per locule[52]. The characteristics of impatiens were valgus lip with Single leaves, spirally arranged, opposite or whorled, stalked or sessile. Sepals 3, sparsely 5, lateral sepals free or connate, entire or toothed, the pistil is composed of 4 or 5 carpels; ovary upper, 4 or 5 compartments, each with 2 to many anatropous ovules[53]. The morphology of *I. guizhouensis* was consistent with the genus of *H. triflora*, although *I. guizhouensis belongs to the genus impatiens*.Here we also found that I. stenosepala and I. linearisepala were four-carpellate[54].

Previous phylogenetic analyses based on plastids, combined plastids and nuclear, or combined plastids and pollen data, which support the monophyly of Balsaminaceae which is consistent with our results of I. balsamina, I. chlorosepala, and I. walleriana belonged to uniflorae. Also, *I. stenosepala* was closer related to *I. linearisepala* than to *I.piufanensis*. So, from the results, using the complete chloroplast genome may be a suitable model tool for understanding the mechanisms of evolution and substantially raises species' discriminatory power in evolutionarily young lineages[55].

## Conclusions

In this study, 12 complete cp genome of Balsaminaceae were analyzed, including six firstly sequenced cp genomes (*I. chlorosepala, I. fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis*, and *I. stenosepala*) comparatively analyzed with

other species. The genus *Impatiens* species of chloroplast genome basic structure and function were similar to the monospecific sister *H. triflora* investigated. Similarly, most divergent genes were detected among these cp genomes, especially in *psbK-psbI, trnT-UGU-trnL-UAA, trnT-GGU-psbD, rpl36-rps8, trnQ-UUG, rpoB-trnC-GCA, trnP-UGG-psaJ, trnK-UUU-rps16, and ycf4-cemA*, which are highly variable regions that can be used as potential markers for species identification and phylogeny. Additionally, based on the ML and BI phylogenomic trees, the trees highly supported the clade of the Balsaminaceae species formed a monophyletic subclade, especially the clusters of *H. triflora* and other impatiens species. This study affirms a useful attempt to understand the phylogeny of this genus and the taxonomic relationship between species and to provide a valuable resource for the study of the genomic evolution of Balsaminaceae.

# Methods

### Materials and DNA Extraction

In total, 12 individuals of Balsaminaceae species were included (Supplementary Table S1), Prof. Haiquan Huang collected and identified all newly sequenced plants in the karst area of Guizhou, additional 6 species were downloaded from GenBank. All voucher specimens were deposited in the plant Laboratory of Southwest Forestry University, Kunming, Yunnan, China(Table 1). Fresh leaves were collected and immediately stored in liquid nitrogen. We extracted the genomic DNA by using the Tiangen DNA Reagent Extraction Kit(Tiangen Biotech, Beijing, China)[56]. And approximately 5-10 μg of genomic DNA was checked using spectrophotometry and their integrity was examined by electrophoresis in 1.5% agarose gel[57].

### Assembly, Annotation, and Analysis

Using an Illumina MiSeq sequencer, purified genomic DNA was sequenced. The contigs were assembled with PE150 reads using SPAdes 3.6.1. The clean data were assembled and manually corrected using GetOrganelle version 1.6.2.[58]. Each assembled cp genome was annotated with GenSeq and DOGAM(Dual Organellar Genome Annotator), and the start and stop codon positions were further searched by the homologous gene identification[59]. Besides, the position of intron-exon junctions of the genes coding proteins, ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs) was confirmed with BLASTN and tRNAscan v1.23 programs[60]. The notes were manually corrected when necessary and verified using GENEROUS R8.0.2 by realigning with references[61]. The physical circular cp genome maps were generated by the OGDrawV1.2 software. Protein-coding genes were compared by comparing the reference species *I.glandulifera* (GenBank KF986530.1) and *I. piufanensis* (GenBank KF986530.1). The GC content were calculated by the GENEROUS R8.0.2 software[62].

### Repeat Sequence Analysis and Simple Sequence Repeats (SSRs) Analysis

The size and location of repeat sequences(forward, palindromic, reverse, and complement repeats) were identified by the REPuter[63]. The sequence identity was 90%, a hamming distance was 3, the minimum repeat size was 30 bp. The online MISA software was used to detect SSRs with the minimum repeat number setting 10, 5, 4, 4, 4, and 4 for mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides, respectively.

### Codon Usage Analysis and Genome Alignment

The software CodonW investigated the distribution of codon usage. The distribution of codon usage was investigated with the RSCU ratio[64]. To detect the divergence hotspots, the online software MAFFT aligned the whole chloroplast genomes[65]. The whole-genome alignment of *Impatiens* and other species was compared by mVISTAprogram in Shuffle-LAGAN mode[66]. DnaSP v5.10 calculated the nucleotide divergence values by using the sliding window method with a window length of 800 bp and a 200 bp step size[67]. Genome-Wide Comparison was aligned with the *H. triflora* cp genome, using the MAUVE software and MAFFT program[68], respectively.

### Phylogenetic Analyses

These chloroplast genomes from seven families within impatiens included 12 Balsaminaceae species, six Primulaceae species, five Ebenaceae species, four Theaceae species, two Saxifragaceae species, four Actinidiaceae species, and one Styracaceae species as outgroups(Accession number: Supplementary Table S8). The aligned sequences were concatenated by MAFFT version 7.222 in default parameter settings[69].The Maximum likelihood (ML) and Bayesian Inference (BI) were conducted for the topologies. The ML analysis was implemented in RAxML v.8.2.9[70] and IQ-TREE ver. 1.6.1. Based on the Akaike information criterion (AIC), the best-fitting GTR+F+I+G4 substitution model with 1000 bootstrap replicates was for ML analyses[71].The Bayesian inference (BI) tree was implemented in MrBayes version 3.2[72]. Based on the Markov chain Monte Carlo (MCMC) algorithm[73], the best-fitting TVM+F+I substitution model with one million generations with four independent heated chains with sampling after every 1000 generations[74]. Fig Tree ver 1.4.2 visualized the output trees[75].

# Abbreviations

BI
Bayesian Inference; bp:base pairs; ETS:Exter transcribed spacer; Gb:Gigabases; IGR:Intergenic region; IR:Inverted repeat; ITS:Internal transcribed spacer; LSC:Long single copy; LSR:Long sequence repeat; MCMC:Markov chain Monte Carlo; ML:Maximum likelihood; NCBI:National Center for Biotechnology Information; NGS:Next-generation sequencing; PCR:Polymerase chain reaction; PI:Parsimony informative; rRNA:ribosomal RNA; SRA:Sequence Read Archive; SSC:Short single copy; SSR:Simple sequence repeat; tRNA:transfer RNA.

# Declarations


### Acknowledgments

We thank Dan Zong helped to teach the software used for the experiments. Our sincere thanks are also to the anonymous reviewers for their comments and suggestions.


### Authors' contributions

C.L. designed the experiment and wrote the manuscript. W.H, Y.L. and X.L. contributed to the sampling. H.S. and H.Y. analyzed the data. B.Y, Q.W. and Y.W. experimented. M.H. and H.H. proofed the manuscript. All authors have read and agreed to the published version of the manuscript.


### Funding

This work was carried out with the support of the National Natural Science Foundation of China [32060364;32060366;31860230], Key Research and Development Plan Program of Yunnan Province [2018BB013], Young and Middle-aged Academic and Technical Leadership Training Project of Yunnan [2015HB046; 2018HB024], and Program for Innovative Research Team (in Science and Technology) in University of Yunnan Province.


### Availability of data and materials

All data generated or analyzed during this study are included in this published article and the complete chloroplast genome sequence of impatiens is deposited in the Genbank. The accession numbers corresponding to the additional datasets used and analyzed in this study can be found in Supplementary Table S8. These were retrieved from the National Center for Biotechnology Information database.

### Ethics approval and consent to participate

Not applicable. The plant was collected in non protected area; no legal authorization/license is required.

### Consent for publication

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Author Details**

[1] College of Landscape Architecture and Horticultural Science, Southwest Forestry University, Kunming, China;

[2] Research and Development Center of Landscape Plants and Horticulture Flowers, Southwest Forestry University, Kunming, China;

[3] Engineering Research Center of Functional Flower Resources and Industrialization in Yunnan Province, Kunming, China;

[4] Department of Landscape Architecture and Plant Science, University of Connecticut, Storrs, CT, 06269, USA;

**Corresponding authors email address**

Name: Haiquan Huang

E-mail: haiquanl@163.com

Name: Meijua Huang

E-mail: xmhhq2001@163.com

# References

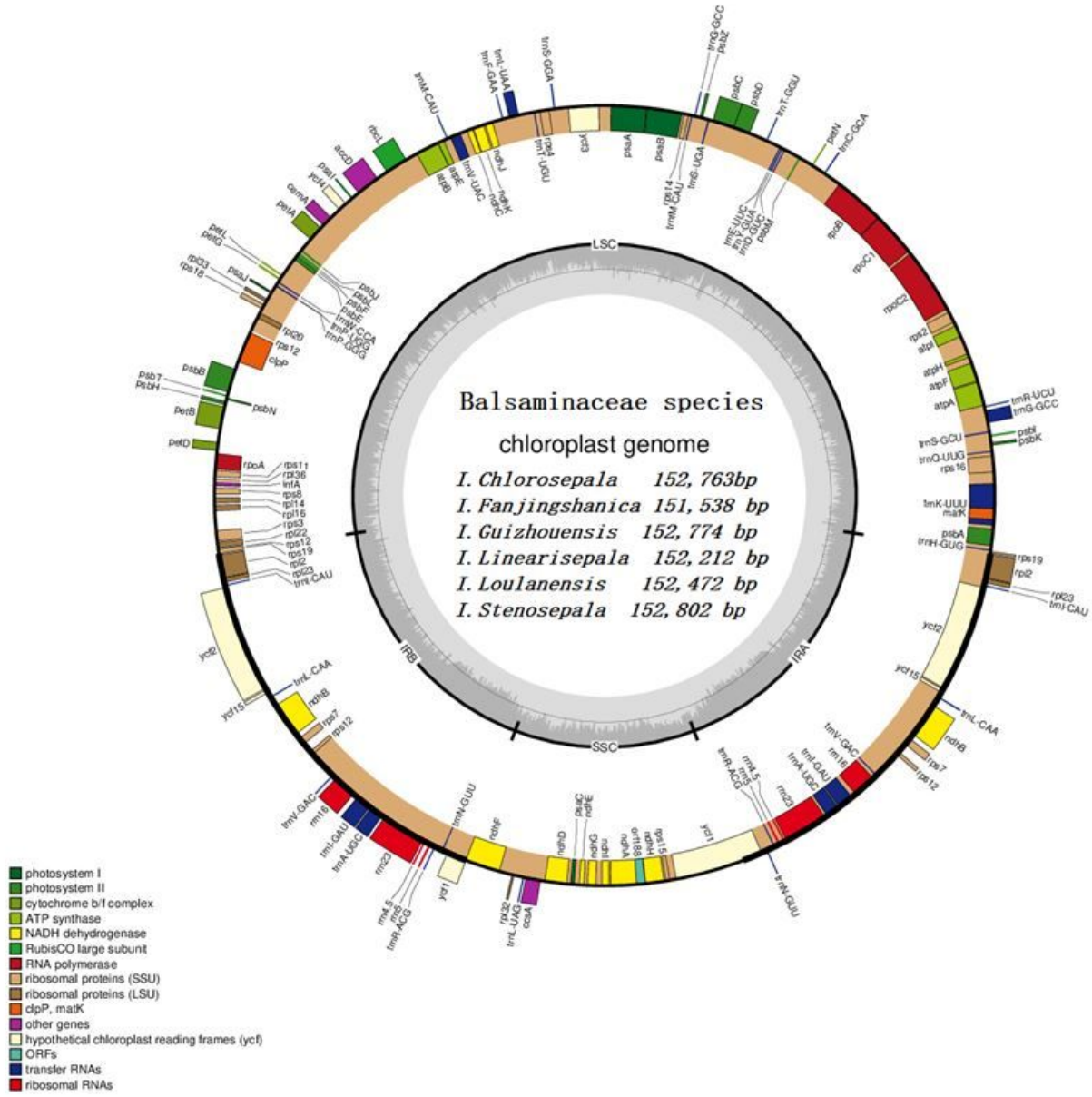
1. Liang H, Zhang Y, Deng J, Gao G, Ding C, Zhang L, Yang R. The Complete Chloroplast Genome Sequences of 14 Curcuma Species: Insights Into Genome Evolution and Phylogenetic Relationships Within Zingiberales. Front. Genet. 2020;11:802.
2. Lee SR, Kim K, Lee BY. et al. Complete chloroplast genomes of all six Hosta species occurring in Korea: molecular structures, comparative, and phylogenetic analyses. BMC Genomics.2019;20,833.
3. Zhou T, Zhu H, Wang J. Complete chloroplast genome sequence determination of Rheum species and comparative chloroplast genomics for the members of Rumiceae. Plant Cell Rep. 2020;39,811-824.
4. Han T, Li M, Li J. et al. Comparison of chloroplast genomes of Gynura species: sequence variation, genome rearrangement and divergence studies. BMC Genomics.2019, 20,791.
5. Ren T, Yang Y, Zhou T, Liu ZL. Comparative Plastid Genomes of Primula Species: Sequence Divergence and Phylogenetic Relationships. Int. J. Mol. Sci. 2018;19,1050.
6. Yang Z, Zhao T, Ma Q, Liang L, Wang G. Comparative Genomics and Phylogenetic Analysis Revealed the Chloroplast Genome Variation and Interspecific Relationships of Corylus (Betulaceae) Species. Front. Plant Sci. 2018;9:927.
7. Li C, Zhao Y, Xu Z, Yang G, Peng J, Peng X. Initial Characterization of the Chloroplast Genome of Vicia sepium, an Important Wild Resource Plant, and Related Inferences About Its Evolution. Front. Genet. 2020;11:73.
8. Liang H, Zhang Y, Deng J, Gao G, Ding C, Zhang L, Yang R. The Complete Chloroplast Genome Sequences of 14 Curcuma Species: Insights Into Genome Evolution and Phylogenetic Relationships Within Zingiberales. Front. Genet. 2020;11:802.
9. Cheng Y, Zhang L, Qi J, Zhang L. Complete Chloroplast Genome Sequence of Hibiscus cannabinus and Comparative Analysis of the Malvaceae Family. Front. Genet. 2020; 11:227.
10. 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.

11. Cho MS, Yang JY, Yang TJ, Kim SC. Evolutionary Comparison of the Chloroplast Genome in the Woody Sonchus Alliance (Asteraceae) on the Canary Islands. Genes. 2019;10,217.
12. Grey-Wilson C. Impatiens in Papuasia. Studies in Balsaminaceae: I. Kew. Bull. 1980b;34, 661-688.
13. Grey-Wilson C. A revision of Sumatran Impatiens. Studies in Balsaminaceae: VIII. Kew. Bull.1989;44,67-105.
14. Janssens SB, Wilson SY, Yuan YM, Nagels A, Smets EF. Huysmans S. A total evidence approach using palynological characters to infer the complex evolutionary history of the Asian Impatiens (Balsaminaceae). Taxon. 2012;61, 355-367.
15. Janssens SB, Knox EB. Huysmans S. Smets EF. Merckx VFST. Rapid radiation of Impatiens (Balsaminaceae) during Pliocene and Pleistocene: result of a global climate change. Mol. Phylogenet. Evol. 2009; 52, 806-824.
16. Grey Wilson C. Impatiens in Papuasia. Studies in Balsaminaceae: I. Kew Bull. 1980;34:661-688.
17. Chen YL. Balsaminaceae. In: Flora Reipublicae Popularis Sinica, Vol. 47. Science Press, Beijing, 2001; pp. 1-243.
18. Cai XZ, Yi RY, Zhuang YH, Cong YY. Kuang RP. Liu KM. Seed coat micromorphology characteristics of Impatiens L. and its systematic significance. Acta. Hort. Sin. 2013;40, 1337-1348.
19. Jiang HF, Zhuang ZH, Hou BW, Shi BJ, Shu HJ, Chen L, Shi GX, Zhang WM. 2017 . Adverse effects of hydroalcoholic extracts and the major components in the stems of Impatiens balsamina L. on Caenorhabditis elegans. Evid Based Complement Alternat Med. 2017;4245830.
20. Kim CS, Bae M, Oh J, Subedi L, Suh WS, Choi SZ. Anti-neurodegenerative biflavonoid glycosides from Impatiens balsamina. J. Nat. Prod.2017; 80, 471-478.
21. Lai HY, Cai MC. Effects of extended growth periods on subcellular distribution, chemical forms, and the translocation of cadmium in Impatienswalleriana, International Journal of Phytoremediation,2016;18:3, 228-234.
22. Ruchisansakun S, Niet T, Van Der T, Janssens SB, Triboun P, Jenjittikul T, Suksathan P. Phylogenetic analyses of molecular data and reconstruction of morphological character evolution in Asian Impatiens section Semeiocardium (Balsaminaceae). Syst. Bot. 2015;40,1063-1074.
23. Rahelivololona EM, Fischer E, Janssens SB, Razafimandimbison SG.Phylogeny, infrageneric classification and species delimitation in the Malagasy Impatiens (Balsaminaceae) PhytoKeys 2018;110: 51-67.
24. Shajitha PP, Dhanesh NR, Ebin PJ, Joseph L, Devassy A, John R, Augustine J, Mathew L. Molecular phylogeny of Balsams (Genus Impatiens) based on ITS regions of nuclear ribosomal DNA implies two colonization events in south India. J. Appl. Biol. Biot. 2016;4, 1-9.
25. Janssens SB, Geuten K, Yuan YM, Song Y, Kupfer P, Smets E. Phylogenetics of Impatiens and Hydrocera (Balsaminaceae) using chloroplast atpB-rbcL spacer sequences. Syst. Bot. 2006; 31, 171-180.
26. Zhang JG, Zhang LB. Impatiens shimianensis sp. Nov (Balsaminaceae): a new species from Sichuan, China, based on morphological and molecular evidence. Syst. Bot. 36, 2011;721-729.
27. Shajitha PP. A combined chloroplast atpB-rbcL and trnL-F phylogeny unveils the ancestry of balsams (Impatiens spp.) in the Western Ghats of India. 3 Biotech 2016;6: 258.
28. Yuan YM, Song Y, Geuten K, Rahelivololona E, Wohlhauser S, Fischer E, Smets E, Küpfer P. Phylogeny and biogeography of Balsaminaceae inferred from ITS sequences. Taxon.2004; 53(2): 391-403.
29. Yuan YM, Song Y, Geuten K, Rahelivololona E, Wohlhauser S, Fischer E, Smets E, Kupfer P. Phylogeny and biogeography of Balsaminaceae inferred from ITS sequence data. Taxon 53, 391-403.
30. Xie DF, Yu Y, Deng YQ, Li J, Liu HY, Zhou SD, He XJ. Comparative Analysis of the Chloroplast Genomes of the Chinese Endemic Genus Urophysa and Their Contribution to Chloroplast Phylogeny and Adaptive Evolution. Int. J. Mol. Sci. 2018, 19, 1847.
31. Zhu ZL, Shi C, Cai NH, Ci XT, Peng JY, Duan AA, Wang DW. The complete chloroplast genome of Yunnanopilia longistaminea (Opiliaceae), an endemic species in southwest China, Mitochondrial DNA Part B, 2019;4:2, 3624-3625.
32. Yan M, Zhao X, Zhou J, Huo Y, Din 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.

33. Cheng H, Li J, Zhang H, Cai B, Gao Z, Qiao Y, Mi L. The complete chloroplast genome sequence of strawberry (Fragaria ananassa Duch.) and comparison with related species of Rosaceae. PeerJ.2017; 5:e3919.
34. Saina JK, Li ZZ, Gichira AW, Liao YY. The Complete Chloroplast Genome Sequence of Tree of Heaven (Ailanthus altissima (Mill.) (Sapindales: Simaroubaceae), an Important Pantropical Tree. Int. J. Mol. Sci. 2018; 19, 929.
35. Luo C, Huang WL, Zhu JP, Feng ZX, Liu YL, Li Y, Li XY, Huang HQ, Huang,MJ. The complete chloroplast genome of Impatiensuliginosa Franch., an endemic species in Southwest China, Mitochondrial DNA Part B. 2019; 4:2, 3846-3847.
36. Cheng Y, Zhang L, Qi J, Zhang L. Complete Chloroplast Genome Sequence of Hibiscus cannabinus and Comparative Analysis of the Malvaceae Family. Front. Genet. 2020; 11:227.
37. Gu C, Tembrock LR, Zheng S, Wu Z. The Complete Chloroplast Genome of Catha edulis: A Comparative Analysis of Genome Features with Related Species. Int. J. Mol. Sci. 2018; 19, 525.
38. Park M, Park H, Lee H, Lee BH, Lee J. The Complete Plastome Sequence of an Antarctic Bryophyte Sanionia uncinata (Hedw.) Loeske. Int. J. Mol. Sci. 2018; 19, 709.
39. Mader M, Pakull B, Blanc-Jolivet C, Paulini-Drewes M, Bouda ZH.N, Degen B, Small I, Kersten B. Complete Chloroplast Genome Sequences of Four Meliaceae Species and Comparative Analyses. Int. J. Mol. Sci. 2018; 19, 701.
40. Ronquist F, Teslenk, M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard .A. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol. 2012; 61, 539-542.
41. 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, 1281-1295.
42. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol. Biol. Evol. 2017;34, 3299-3302.
43. Campos V, Lessa SS, Ramos RL, Shinzato MC, Medeiros TAM. Disturbance response indicators of Impatienswalleriana exposed to benzene and chromium, International Journal of Phytoremediation, 2017;19:8, 709-717.
44. Li W, Zhang C, Guo X, Liu Q, Wang K. Complete chloroplast genome of Camellia japonica genome structures, comparative and phylogenetic analysis. PLoS ONE.2019; 14(5): e0216645.
45. Asaf S, Khan AL, Aaqil Khan M, Muhammad Imran Q, Kang SM, Al-Hosni K, et al. Comparative analysis of complete plastid genomes from wild soybean (Glycine soja) and nine other Glycine species. PLoS ONE.2017; 12(8): e0182281.
46. Chen YL. Notulae de genere Impatiens L. flora Sinicae. Acta Phytotax. Sin. 1978;16, 36-55.
47. Fujihashi H, Akiyama S, Ohba H. Origin and relationships of the Sino-Himalayan Impatiens (Balsaminaceae) based on molecular phylogenetic analysis, chromosome numbers and gross morphology. J. Jap. Bot. 2002;77, 284-295.
48. Yuan Y, Song Y, Geuten K, Rahelivololona E, Fischer E, Smets E, Küpfer P. Phylogeny and biogeography of Balsaminaceae inferred from ITS sequences. Taxon. 2004; 53, 391-403.
49. Cafa G, Baroncelli R, Elliso, CA, Kurose D. Impatiens glandulifera (Himalayan balsam) chloroplast genome sequence as a promising target for populations studies. PeerJ. 2020; 8:e8739.
50. Yu SX, Janssens SB, Zhu XY, Lid, en M, Gao, TG, Wang W. Phylogeny of Impatiens (balsaminaceae): integrating molecular and morphological evidence into a new classification. Cladistics. 2016; 32(2): 179-197.
51. Tamboli AS, Dalavi JV, Patil SM, Yadav SR, Govindwar SP. Implication of ITS phylogeny for biogeographic analysis, and comparative study of morphological and molecular interspecies diversity in Indian Impatiens. Meta Gene. 2018;16, 108-116.
52. Chen YL. Balsaminaceae. In: Flora Reipublicae Popularis Sinica, Vol. 47. Science Press, Beijing. 2001; pp. 1-243.
53. Yu SX. Balsaminaceae of China. Peking University Press, Beijing. 2012.
54. Li Q, Zhang XS, Cao JQ, Guo ZH, Lou YT, Ding M, Zhao YQ. Depside derivatives with anti-hepatic fibrosis and anti-diabetic activities from Impatiens balsamina L. flowers. Fitoterapia. 2015; 105:234-239.
55. Dong W. Xu C. Wen J. Evolutionary directions of single nucleotide substitutions and structural mutations in the chloroplast genomes of the family Calycanthaceae. BMC Evol Biol. 2020;20,96.

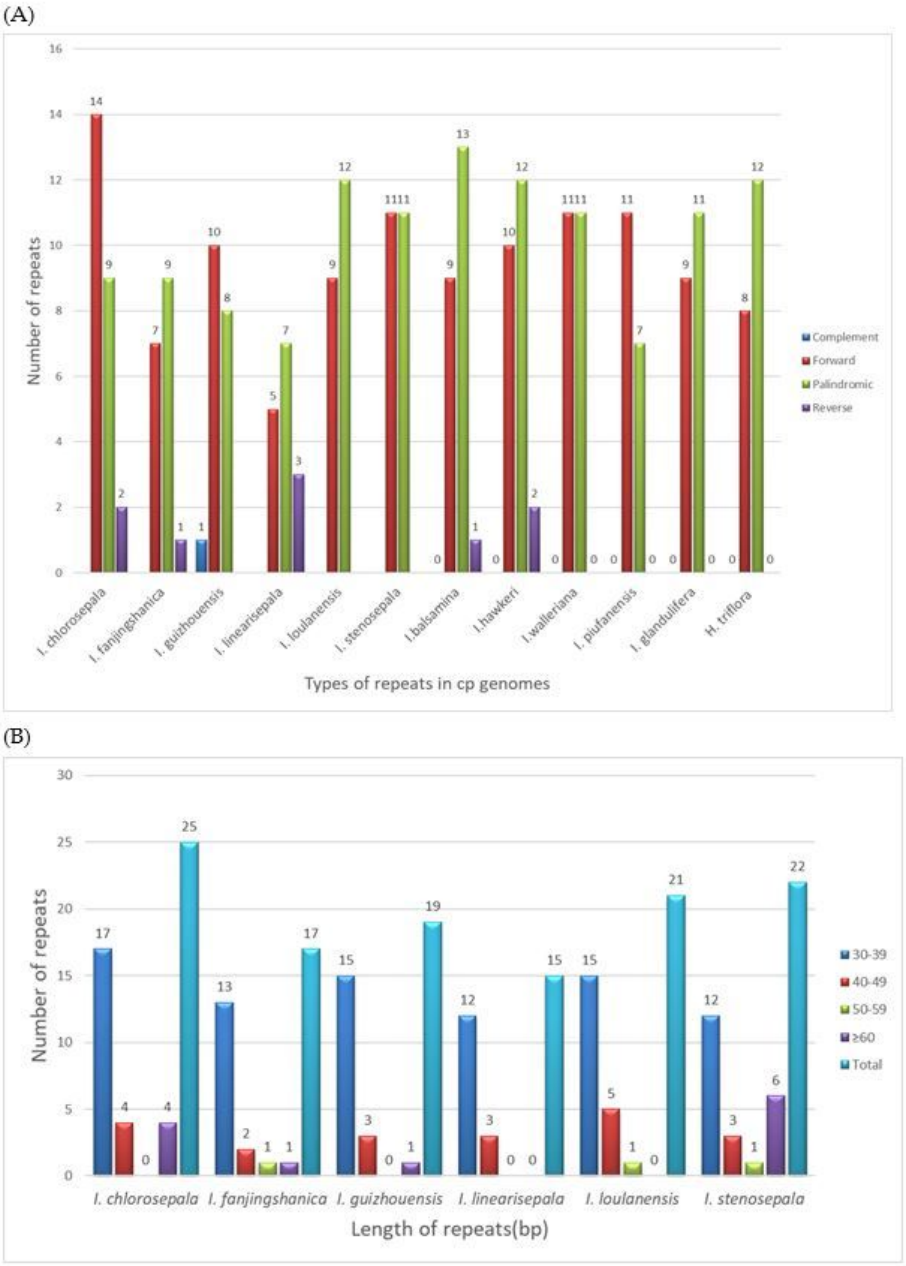
56. Jin HP, Lee JJ. The complete plastid genome of Scopolia parviflora (Dunn.) Nakai(Solanaceae) Korean J. Pl. Taxon. 2016; 46(1): 60-64.
57. Zhu, S.Y, Niu, Z.T, Yan, W.J, Xue, Q.Y, Ding, X.Y. The complete chloroplast genome sequence of Anoectochilusemeiensis, Mitochondrial DNA Part A. 2016, 27:5, 3565-3566.
58. Langmead, B, Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012, 9, 357-359.
59. Tillich, M, Lehwark, P, Pellizzer, T, Ulbricht-Jones, E.S, Fischer, A, Bock, R, Greiner, S. GeSeq—Versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017, 4, W6-W11.
60. Peter S, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33, 686-689.
61. Huang Y, Yang Z, Huang S, An W, Li J, Zheng X. Comprehensive Analysis of Rhodomyrtus tomentosa Chloroplast Genome. Plants. 2019; 8, 89.
62. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings Bioinform. 2008;9, 299-306.
63. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: A web server for microsatellite prediction. Bioinformatics. 2017; 33, 2583-2585.
64. Katoh K. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30, 3059-3066.
65. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 2019;20, 1160-1166.
66. Beier S, Thomas T, Münch T, Scholz U, Mascher M. MISA-web: A web server for microsatellite prediction. Bioinformatics. 2017; 33, 2583-2585.
67. Rozas J, Ferrer-Mata, A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol. Biol. Evol. 2017; 34, 3299-3302.
68. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics 2010; 26, 1899-1900.
69. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics. 2010; 26, 1899-1900.
70. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 2015; 32, 268-274.
71. Posada D. jModelTest: Phylogenetic Model Averaging. Mol. Boil. Evol. 2008,25,1253-1256.
72. Swofford DL. Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5. 2001.
73. Gascuel O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol. Boil. Evol. 1997,14, 685-695.
74. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings Bioinform. 2008; 9, 299-306.
75. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol. 2012; 61, 539-542.

## Figures



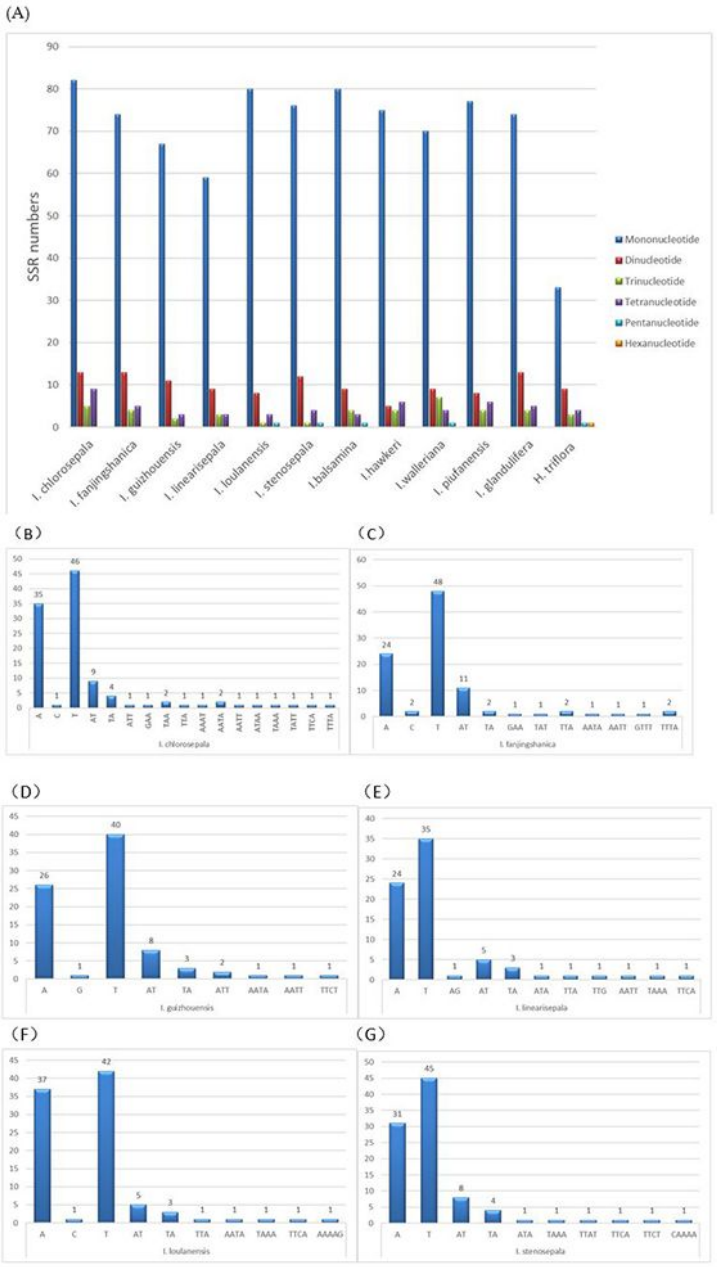

## Figure 1

Chloroplast genome structure of Impatiens species(I. chlorosepala, I.fanjingshanica, I. guizhouensis, I. linearisepala, I. loulanensis, and I. stenosepala.). Genes shown outside the map circles are transcribed clockwise, while those drawn inside are transcribed counterclockwise. Genes from different functional groups are color-coded according to the key at the top right. The positions of long single copy (LSC), short single copy (SSC), and two inverted repeats (IR: IRA and IRB) regions are shown in the inner circles.



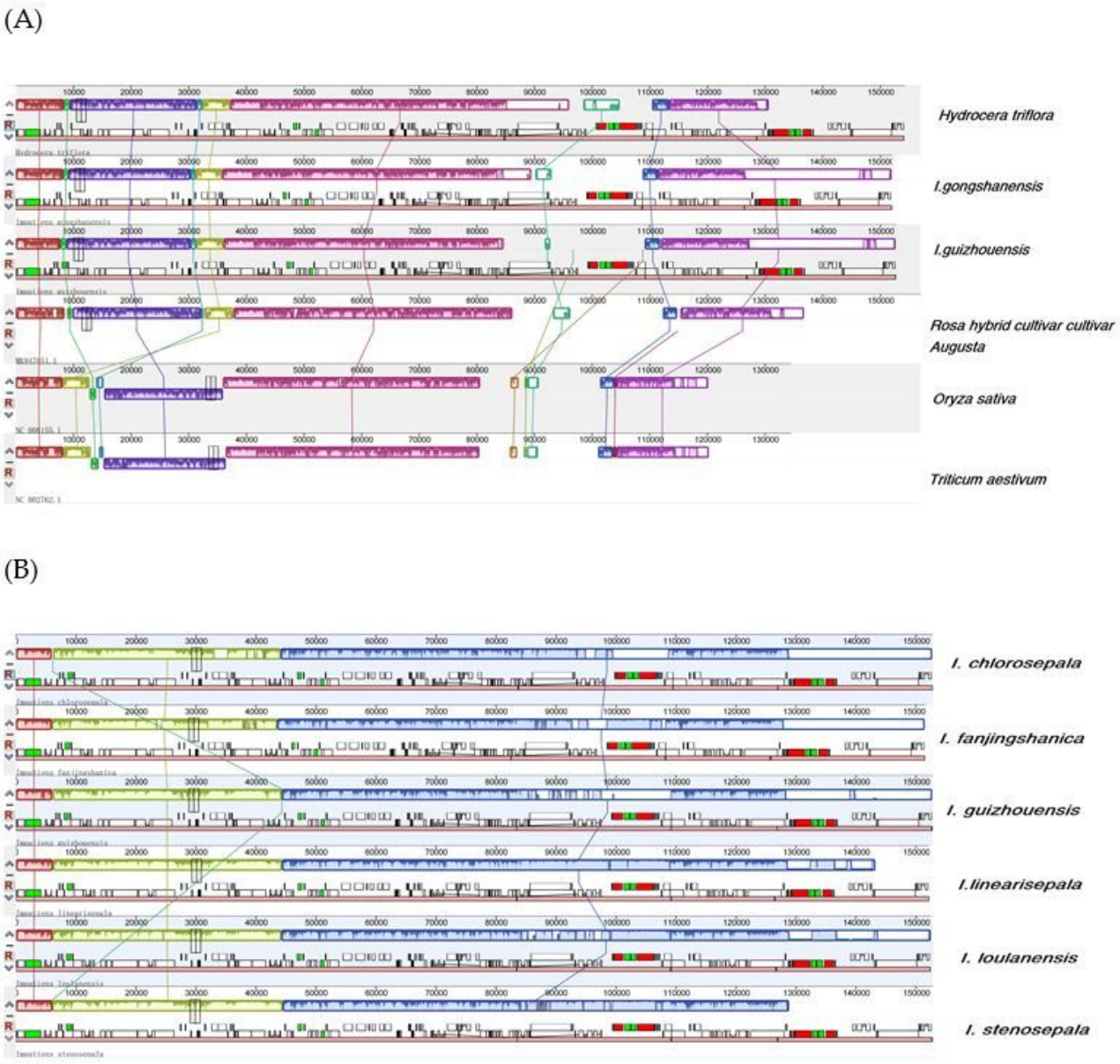

## Figure 2

Repeated sequences in Balsaminaceae chloroplast genomes. (A) Total of four repeat types in twelve Balsaminaceae chloroplast genomes; (B) Numbers of repeat sequences by length



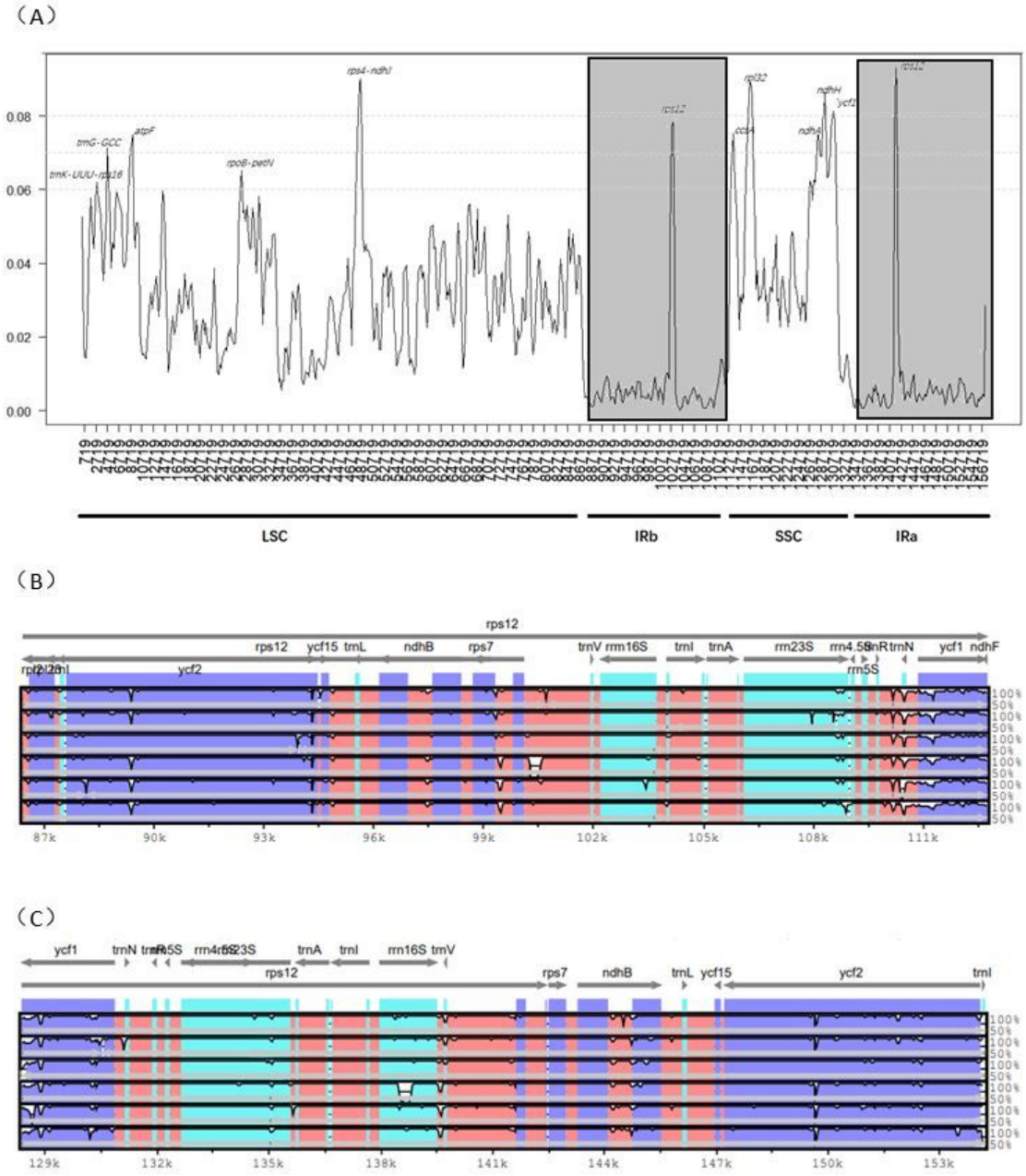

## Figure 3

SSR loci analysis of twelve Balsaminaceae chloroplast genomes. (A) Numbers of different SSRs types detected in the twelve genomes; (B-G): Frequency rates of identified SSR motifs in different repeat class types



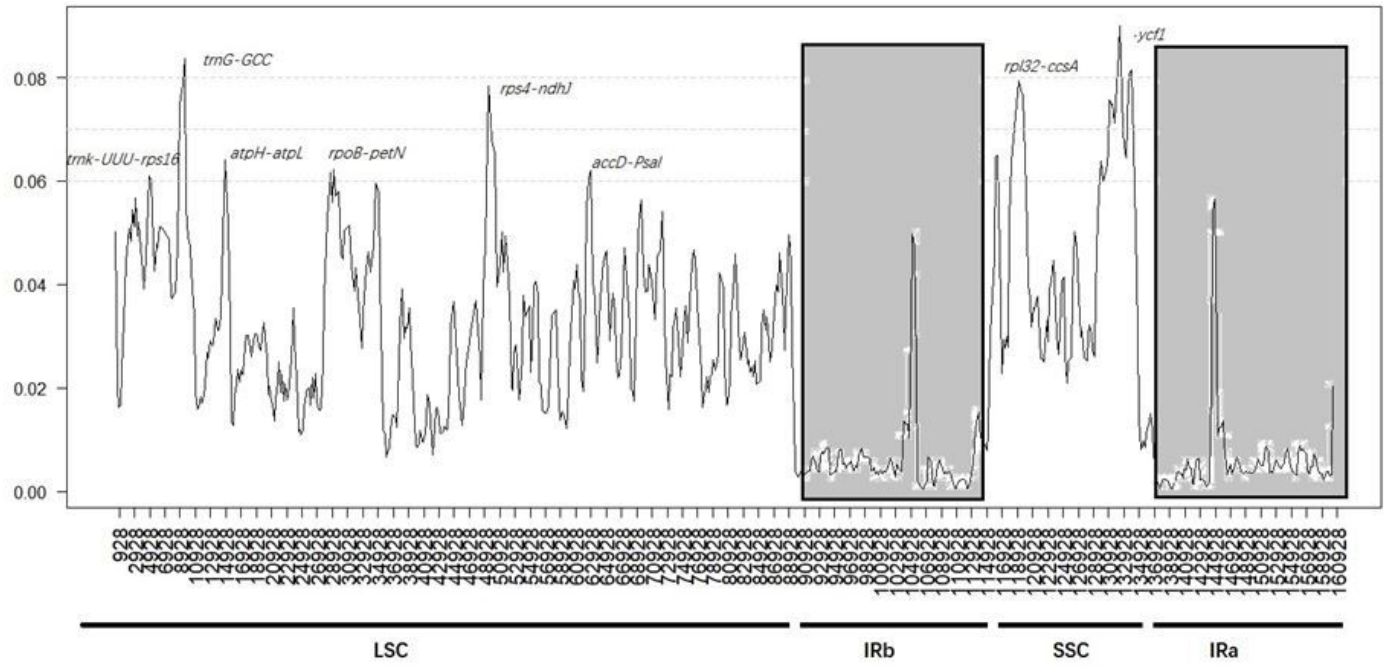

## Figure 4

Mauve alignment. (A)Two rearrangements concerning the dicot plastome with LSC and IRB intermolecular recombination (B) Mauve alignment of six Balsaminaceae plastomes revealing no interspecific rearrangement.



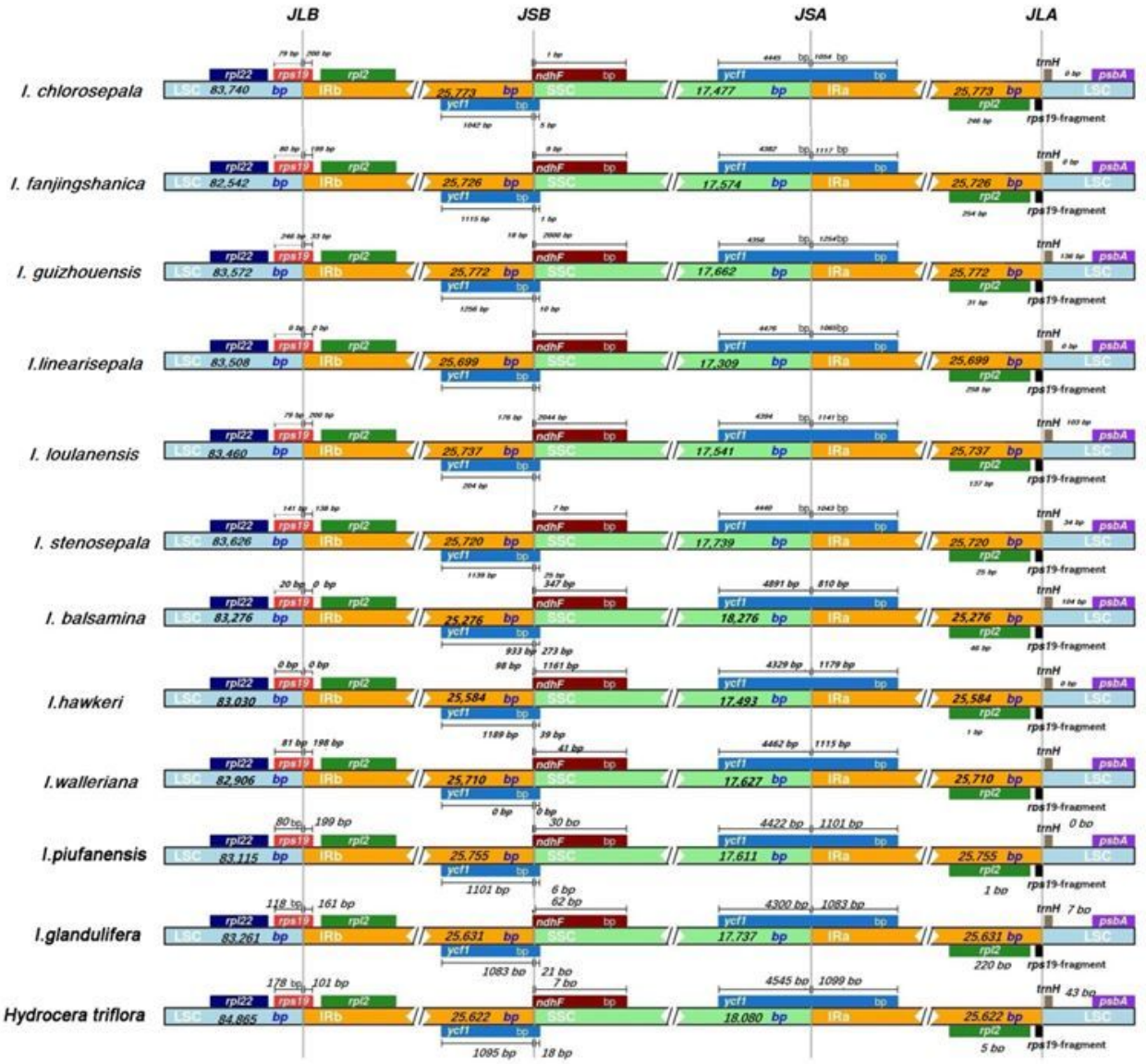

## Figure 5

(A)Sliding window analysis based on the newly sequenced cp genomes of Balsaminaceae species. (B) The sequence divergence from 87,000 bp to 111,000 bp is visualized by the mVISTA program. The vertical scale indicates percentage identity, ranging from 50 to 100%. (C) The sequence divergence from 129,000 bp to 153,000 bp is visualized by the mVISTA program.




## Figure 6

Sliding window analysis based on the cp genomes of 12 Balsaminaceae species. Window length: 2000 bp; step size: 200 bp. X-axis: the position of the midpoint of a window. Y-axis: nucleotide diversity of each window.




## Figure 7

Comparison of the borders of large single copy (LSC), small single copy (SSC), an inverted repeat (IR) regions among 12 cp genomes. The number above the gene features means the distance between the ends of genes and the border sites.

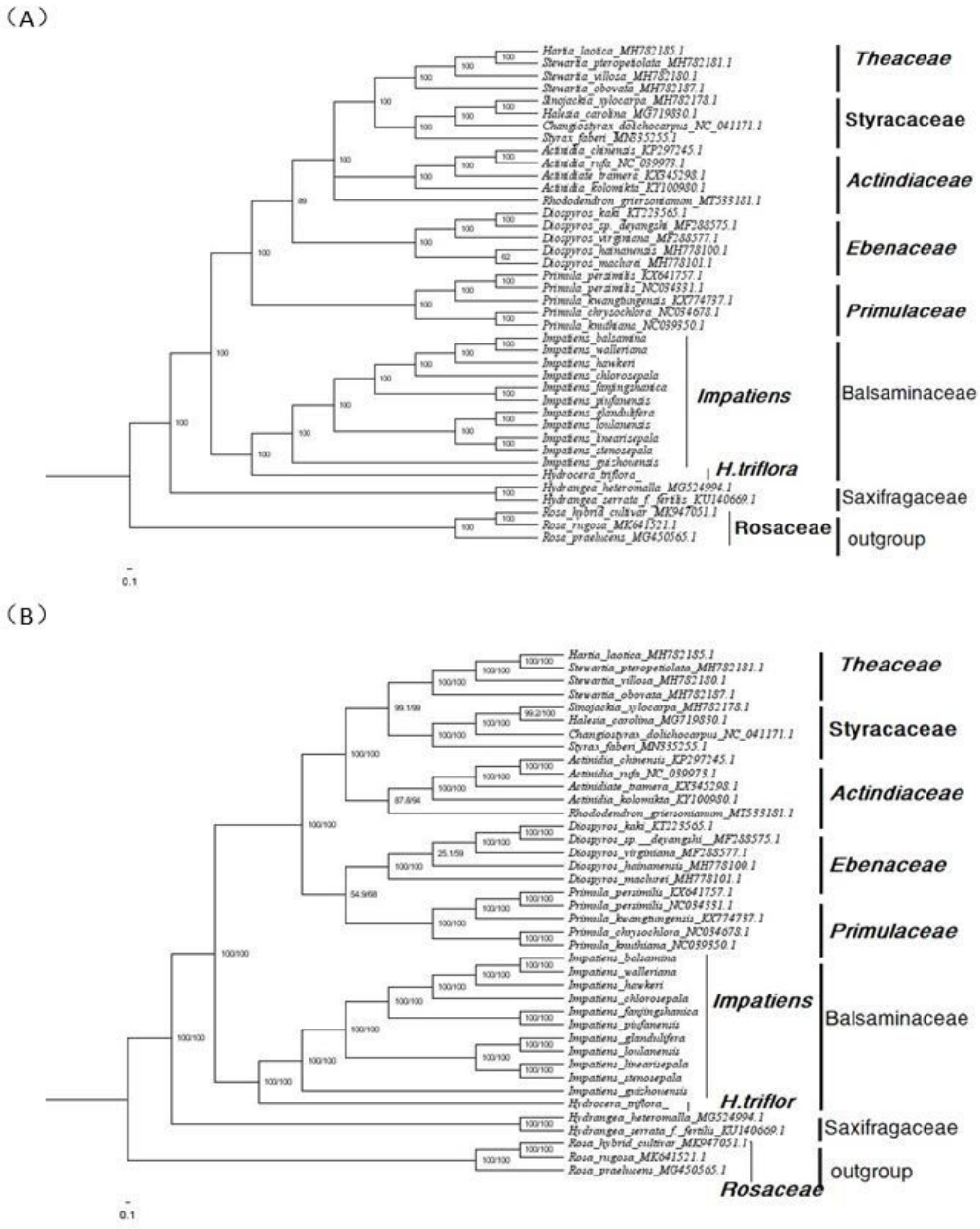



**Figure 8**

Phylogenetic tree of Impatiens species within the Balsaminaceae. The entire genome data set was analyzed using maximum likelihood (ML) and Bayesian information (BI). The numbers above and below the branches represent bootstrap values in the ML and BI trees. The green color represents the positions of Impatiens species. (A) ML tree; (B) BI tree.

## Supplementary Files

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

- Additionalfile1.zip
- Additionalfile2.zip